JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 8 (2020), Issue 6   ISSN 2291-9694   Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Corrigenda and Addenda

Original Paper

Using Electronic Data Collection Platforms to Assess Complementary and Integrative Health Patient-Reported Outcomes: Feasibility Project

Jolie N Haun[1,2], EdS, PhD; Amy C Alman[1,3], PhD; Christine Melillo[1], PhD, BSN; Maisha Standifer[1,4], PhD; Julie McMahon-Grenz[1], MS, OTR; Marlena Shin[5], MPH, JD; W A Lapcevic[1], MS, MPH; Nitin Patel[6], MPH; A Rani Elwy[7,8], PhD

[1]Research Service, James A. Haley VA Medical Center, Tampa, FL, United States
[2]Department of Community & Family Health, College of Public Health, University of South Florida, Tampa, FL, United States
[3]Department of Public Health, University of South Florida, Tampa, FL, United States
[4]Department of Pharmacy Practice, College of Pharmacy, University of South Florida, Tampa, FL, United States
[5]Center for Healthcare Organization and Implementation Research, Veterans Affairs Boston Healthcare System, Boston, MA, United States
[6]Performance Improvement and Reporting, VHA Office of Community Care, Department of Veteran Affairs, Washington, DC, United States
[7]Center for Healthcare Organization and Implementation Research, Bedford Veterans Affairs Medical Center, Bedford, MA, United States
[8]Brown University, Providence, RI, United States

Corresponding Author:
Jolie N Haun, EdS, PhD
Research Service
James A. Haley VA Medical Center
8900 Grand Oak Circle
Tampa, FL, 33637
United States
Phone: 1 813 558 7622
Email: Jolie.Haun@va.gov

Abstract

Background: The Veteran Administration (VA) Office of Patient-Centered Care and Cultural Transformation is invested in improving veteran health through a whole-person approach while taking advantage of the electronic resources suite available through the VA. Currently, there is no standardized process to collect and integrate electronic patient-reported outcomes (ePROs) of complementary and integrative health (CIH) into clinical care using a web-based survey platform. This quality improvement project enrolled veterans attending CIH appointments within a VA facility and used web-based technologies to collect ePROs.

Objective: This study aimed to (1) determine a practical process for collecting ePROs using patient email services and a web-based survey platform and (2) conduct analyses of survey data using repeated measures to estimate the effects of CIH on patient outcomes.

Methods: In total, 100 veterans from one VA facility, comprising 11 cohorts, agreed to participate. The VA patient email services (Secure Messaging) were used to manually send links to a 16-item web-based survey stored on a secure web-based survey storage platform (Qualtrics). Each survey included questions about patient outcomes from CIH programs. Each cohort was sent survey links via Secure Messaging (SM) at 6 time points: weeks 1 through 4, week 8, and week 12. Process evaluation interviews were conducted with five primary care providers to assess barriers and facilitators to using the patient-reported outcome survey in usual care.

Results: This quality improvement project demonstrated the usability of SM and Qualtrics for ePRO collection. However, SM for ePROs was labor intensive for providers. Descriptive statistics on health competence (2-item Perceived Health Competence Scale), physical and mental health (Patient-Reported Outcomes Measurement Information System Global-10), and stress (4-item Perceived Stress Scale) indicated that scores did not significantly change over time. Survey response rates varied (18/100, 18.0%-42/100, 42.0%) across each of the 12 weekly survey periods. In total, 74 of 100 participants provided ≥1 survey, and 90% (66/74) were female. The majority, 62% (33/53) of participants, who reported the use of any CIH modality, reported the use of two or more unique modalities. Primary care providers highlighted specific challenges with SM and offered solutions regarding staff involvement in survey implementation.

Conclusions: This quality improvement project informs our understanding of the processes currently available for using SM and web-based data platforms to collect ePROs. The study results indicate that although it is possible to use SM and web-based survey platforms for ePROs, automating scheduled administration will be necessary to reduce provider burden. The lack of significant change in ePROs may be due to standard measures taking a biomedical approach to wellness. Future work should focus on identifying ideal ePRO processes that would include standardized, whole-person measures of wellness.

(JMIR Med Inform 2020;8(6):e15609) doi:10.2196/15609

KEYWORDS

integrative medicine; health information technology; health services research; mobile phone; patient-reported outcomes; veteran

Introduction

Background

The Veterans Health Administration (VHA) is committed to increasing the use of health information technology (HIT) to promote personalized and patient-driven health services, including complementary and integrative health (CIH) [1]. The integration of HIT has the potential to increase quality and access to CIH, enhance patient outcomes, increase efficiency, and decrease costs [2,3]. Effective implementation of integrated HIT, such as telehealth and electronic health records, is a priority for the VHA [4].

The VHA prioritizes access to CIH [5,6]. Complementary treatment is based on Eastern medicine philosophies and includes a variety of modalities, including but not limited to yoga, mindfulness, and acupuncture [7]. Integrative health is the use of both Western or *traditional* treatments in combination with *complementary* treatments [8]. Integrative health tends to lower any power differential between patient and provider as well as focus on contextual health [8]. Recent VHA programs focus on improving access to CIH modalities for all veterans, making integration of CIH into health plans a priority within the VHA [1,6].

Within the VHA, HIT for health care delivery is not only beneficial for providers and their delivery of health care services but is also advantageous for meeting patient-specific needs. Using HIT for health care delivery facilitates timely reporting of outcomes (ie, improves recall accuracy) and eliminates the potential for misplaced documentation. The process empowers patients to make informed health care decisions, improves patient satisfaction, and streamlines organizational processes that, historically, were barriers to the delivery of health care services [9]. Electronic health communication systems have improved clinical effectiveness and enhanced communication between patients and providers [10]. Patients managing chronic conditions have reported satisfaction in web-based reporting systems that facilitate effective communication of biomedical metrics (eg, blood pressure and weight) [11]. Recent studies have demonstrated the usability and implementation of HIT platforms to collect electronic patient-reported outcomes (ePROs), such as biomedical metrics for disease management [12-14].

Patient-centered outcome measures for CIH are increasingly important as the use of nontraditional therapies and treatments increases in clinical settings [15]. Several research studies support the advantages of collecting ePROs through mobile technology application, secure messaging (SM), or text messaging [16-18], but there is little published research on the use of HIT to collect CIH-related ePROs within the unique Veteran Administration (VA) system [19]. There is a need to develop a practical process for collecting and integrating ePROs to improve patient care within the VHA [20-22].

Objectives

This study aimed to (1) determine a practical process for collecting and integrating ePROs using SM and a web-based survey platform and (2) conduct analyses of pilot survey data using repeated measures to estimate the effects of CIH on patient outcomes. This paper provides lessons learned from the implementation of ePRO survey methods within the VA.

Methods

Design

This 1-year pilot project engaged veterans attending CIH appointments within one VA facility. We used process documentation, quantitative repeated measures surveys, and qualitative interviews to meet project objectives.

Veterans Sample

This project used a convenience sample. Project team members reached out to local primary care providers (PCPs; N=21) who were known to make referrals to CIH program services at the project site. These PCPs were identified as early adopters in the CIH program. Early adopter PCPs were asked to provide a list of potential participants who participated in at least one CIH program or activity at the project site based on their personal knowledge of veteran wellness activities.

Of the responsive providers (n=5), a participant pool of 227 veterans was identified. The project site had a robust integrative health program at the time of participant pool identification. Providers often used referral to the integrative health program as a proxy for the inclusion criteria of the CIH program or activity participation. Additionally, at the time of participant pool identification, providers based their list of potential participants on general knowledge of veterans having used CIH program services in the recent past. Between participant pool identification and completion of recruitment, up to four months had passed, generating a subgroup of participants that were no longer participating in CIH programs during the project. Our final convenience sample of 100 veterans was recruited via telephone to participate in the project based on the following inclusion criteria: (1) participated in at least one CIH program or activity at the project site and (2) had access to SM to complete the web-based surveys.

Two electronic messaging platforms for data collection were identified by operational partners and clinical providers: SM, a messaging platform provided to veterans through MyHealtheVet (a veteran-facing health care portal), and ANNIE, a mobile app that allows messaging, alerts, and push notifications, and can send messages to mobile phones that are not smartphones. ANNIE was not accessible for utilization at the time of project implementation. Owing to the unavailability of ANNIE, only the SM web-based platform was used. SM, as a messaging service, requires its own unique set of processes to send individual messages to each veteran in the project. We developed a customized protocol for using SM to collect and integrate ePROs, which was tested in this project.

Participants were grouped into cohorts to facilitate data collection. Each cohort was the result of the project team making telephone calls once per week (typically Friday). Participants who agreed to participate were included in that week's cohort and notified of the intent of the study, number and frequency of assessments, time burden for each assessment, and the need to access their SM account routinely (typically Monday). Thus, a cohort consisted of veterans who received electronic requests for survey completion at the same time points. The 11 cohorts ranged in size from 2 to 21 participants, with more participants in the initial cohorts and fewer in the latter cohorts, with an average cohort size of 9. Process evaluation interviews were conducted with 5 PCPs involved in the SM implementation of the ePRO survey to assess the acceptability, appropriateness, and feasibility of using SM to direct veterans to the web-based ePRO survey in future VA sites.

Data Collection Procedure

The effects of CIH activities may change over time and are often measured at variable time points [23]. As such, all participants within a cohort were sent the survey link via SM at 6 time points: weeks 1 through 4, week 8, and week 12. The time points were selected to evaluate both short- (1-4 weeks) and long-term (week 8 and week 12) assessment of ePRO responses. The initial weekly assessments were used not only to capture an initial change in ePROs (based on new participation in CIH activities) but also to test the use of weekly survey links (eg, how to distinguish survey links and reminders, and how to manage data). More than six time points were considered potentially burdensome for the participants. Rolling enrollment and sending of SM links occurred over a 25-week period. Each participant was sent an SM message with a link and a unique personal identification number at the beginning of each week and again 3 days later to nonrespondents. Each link contained a web-based survey consisting 16 items from the scales described in the *Scales* section below as well as items to indicate the types of CIH programs in which the participant was currently participating. Survey data were collected and managed via Qualtrics), a web-based survey platform.

Process evaluation interviews were conducted via telephone. One researcher conducted the interview, and another researcher took extensive notes on this interview, capturing as much verbatim information as possible. Interviews lasted approximately 30 min and focused on capturing stakeholders' perspectives on the acceptability, appropriateness, and feasibility of using the ePRO survey with veterans receiving CIH services through the SM platform. Our conceptual framework is derived from the categorization of implementation outcomes by Proctor et al [24], the implementation outcomes framework. We specifically assessed (1) acceptability or the extent to which adopting the ePRO survey is agreeable, palatable, or satisfactory among key stakeholders; (2) appropriateness, the perceived fit, relevance, or compatibility of the innovation or evidence-based practice for VA primary care providers and staff; and (3) feasibility or the extent to which the ePRO survey can be successfully used or carried out within VA primary care.

Scales

Participants completed a 16-item web-based survey consisting of items that measured health competence on the 2-item Perceived Health Competence Scale (PHCS-2), physical and mental health on the 10-item Patient Reported Outcome Measurement Information System (PROMIS-10), and perceived stress on the 4-item Perceived Stress Scale 4 (PSS-4). These scales were chosen to obtain data across a broad range of patient-reported health outcomes, in which CIH modalities may have an impact. Survey links did not have an expiration time designated so they could be completed as assigned or at any point during the project. Demographic data were extracted from the electronic medical records of all participants.

The PHCS-2 is a balanced subscale consisting of 2 questions (1 positively worded and 1 negatively worded) chosen from the larger 8-item PHCS and measures the degree to which an individual feels capable of reasonably managing his or her own health outcomes. The PHCS has previously shown to be a valid and reliable measure of health competence [25]. Total scores range from 2 to 10, with higher scores reflecting increased perceived health competence.

The PROMIS-10 short form consists of 10 items that assess the general domains of physical and mental health and functioning. PROMIS was developed and validated by investigators at the National Institutes of Health to provide clinicians and researchers with accessible item banks to measure patient-reported health status [26,27]. Raw scores for the physical and mental domains ranged between 4 and 20. Higher scores reflect better physical and mental functioning.

The PSS-4 measures the degree to which situations in one's life over the past month are appraised as stressful. Items were designed to detect how unpredictable, uncontrollable, and overloaded respondents find their lives. PSS-4 was derived from the longer PSS-14 [28]. As the questions are of a general nature and are not directed at any particular subpopulation, using the abbreviated version (or any version) with a diverse population is predicted to yield equally reliable results. The score ranges from 0 to 16, with higher scores representing more stress.

Data Analysis

Data are presented as mean (SD) for continuous data and frequencies and percentages for categorical data. Chi-square tests and two-tailed *t* tests were used to compare demographics between survey responders and nonresponders. Survey response rates were considered a measure of ePRO collection process feasibility, with a focus on practicality [29]. Survey response rates were defined as the percentage of participants who responded to at least one survey question over the examination period. We also examined the length of time to completion and completeness of responses to the scale items (PHCS-2, PROMIS-10, and PSS-4). Final scores for the scales were calculated and used in the analyses only when all items were complete. A sensitivity analysis of multiple imputation of missing items indicated no change in conclusions, so we retained only the complete survey data for analysis. Linear mixed effects analysis was conducted separately for each PRO measure to analyze within-subject changes in responses over time. This procedure considers the correlation that occurs for repeated measurements and can handle when the number of assessments is unequal between subjects. All models included fixed effects for time, age, gender, race, and Hispanic ethnicity and an indicator of any CIH use as well as a random intercept. The change in each scale over time did not significantly vary between subjects, so a random effect of time was not included in the models. A *P* value <.05 was used to assess significance. Analyses were performed using SAS software, version 9.4.

Interview notes were analyzed using a directed content analysis approach [30], through inductive open coding [31] to identify themes related to our a priori framework of acceptability, appropriateness, and feasibility [24] that arose through the interviews. Two researchers coded each interview after it took place, coding each set of notes independently and then meeting to identify codes and collapse ideas into broader themes. All coding took place using a discussion and consensus approach, where discrepancies in coding were discussed, evidence was presented, and then a consensus on the coding process occurred. This process continued with each interview, and after the fifth interview, when no new ideas emerged, the researchers determined that saturation had been met, and no additional interviews took place.

Results

Overview

Table 1 displays the participant characteristics. The mean age was 54.7 (SD 9.4) years, and the majority were female (66/74, 90%). The majority were white (49/74, 66%) and not Hispanic or Latino (66/74, 89%). Despite disruption in link connectivity access in earlier messages, 74/100 (74%) provided at least one survey response, 53/100 (53%) provided at least two responses, 30/100 (30%) provided at least three responses, 13/100 (13%) provided at least four responses, and 3/100 (3%) provided at least five responses. However, only 2 participants responded to all 6 surveys. Nonresponders were significantly younger than responders (mean 49.0, SD 10.0 years vs 56.7, SD 8.4 years; *P*<.001) and less likely to be white (46.2% vs 66.2%; *P*=.02) but did not differ by gender (female 92.3% vs 89.2%; *P*=.65) or Hispanic ethnicity (12.0% vs 10.8%; *P*=.87).

Response rates were lowest in week 1 (18%) and generally increased in the subsequent weeks (24% for week 2, 34% in week 4, and 33% in week 8). The highest response rate was in the final week of participation, where 42% of participants responded to the survey. This is commensurate with the established expectations of response rates based on our previous work [19,32]. The majority of responses (77.7%) were received within one week with a median of 2 days (25th-75th percentile: 1-5 days) of the SM request to complete the survey.

A total of 74 participants responded to at least one of the survey links sent to them via SM, resulting in 175 responses over the data collection period. Of these, 150 (85.7%) were completed for all 16 items from the 3 scales (PHCS-2, PROMIS-10, and PSS-4). Completeness was lowest for the PROMIS-10 (153/175, 87.4%) and highest for the PHCS-2 (169/175, 96.6%), although the PSS-4 was similar with a completeness of 168/175 (96.0%). A total of 21 surveys were missing between 1 and 14 items across the 3 scales, with the majority missing just 1 item (17/21, 81.0%). There were 2 surveys (1.1%) missing 1 item on the PHCS-2, 16 surveys (9.1%) missing 1 item on the PROMIS-10, and 1 survey (0.57%) missing 1 item on the PSS-4. There was 1 survey missing 2 items on the PROMIS-10 and 4 surveys (2.3%) missing all 16 items from the 3 scales. One additional survey was missing all items for the PROMIS-2, and 2 surveys were missing all items for the PSS-4.

Descriptive statistics of the computed scores from each of the scales are presented in Table 2. Least squares means (SE) for each scale adjusted for age, gender, race, and Hispanic ethnicity are displayed in the table with the corresponding N for each scale (PROMIS-10, PSS-4, PHCS-2; the n values for the mental and physical domains of the PROMIS-10 are the same). In the mixed models, the linear fixed effect of time was not significant for any of the scales, indicating that the scores did not change over time. Age, gender, race, Hispanic ethnicity, and any CIH use were not significant predictors in any of the models.

Table 1. Participant characteristics.

Characteristics	Responders (n=74)	Nonresponders (n=26)	*P* value[a]
Age (years), mean (SD)	56.7 (8.4)	49.0 (10.0)	<.001
Sex, n (%)			
Female	66 (89.2)	24 (92.3)	.65
Race, n (%)			
White	49 (66.2)	12 (46.2)	.02
Black	20 (27.0)	6 (23.1)	—[b]
Native American	1 (1.4)	0 (0.0)	.02
Asian	1 (1.4)	2 (7.7)	.02
Native Hawaiian	1 (1.4)	1 (3.8)	.02
Multiracial	0 (0.0)	2 (7.7)	.02
Unknown	2 (2.7)	3 (11.5)	.02
Ethnicity, n (%)			.87
Not Hispanic or Latino	66 (89.2)	22 (88.0)	
Hispanic or Latino	8 (10.8)	3 (12.0)	
Number of surveys completed, n (%)			—
1	21 (28.4)	N/A[c]	
2	23 (31.1)	N/A	
3	17 (23.0)	N/A	
4	10 (13.5)	N/A	
5	1 (1.4)	N/A	
6	2 (2.7)	N/A	

[a] *P* value from *t* tests, chi square tests, or Fisher exact tests.

[b] Data unavailable.

[c] N/A: not applicable.

Table 2. Least squares means of computed scores for each scale per week.

Scale	Week						*P* value
	1 (n=18)	2 (n=24)	3 (n=23)	4 (n=31)	8 (n=31)	12 (n=42)	
PHCS-2[a], mean (SE)	6.8 (0.36)	6.5 (0.32)	6.6 (0.32)	6.0 (0.28)	6.4 (0.28)	6.6 (0.26)	.22
PROMIS-10[b] mental, mean (SE)	10.3 (0.56)	11.4 (0.51)	11.1 (0.50)	10.2 (0.46)	10.8 (0.47)	10.4 (0.44)	.09
PROMIS-10 physical, mean (SE)	12.0 (0.47)	11.2 (0.45)	11.3 (0.44)	11.2 (0.40)	11.8 (0.41)	11.3 (0.39)	.17
PSS-4[c], mean (SE)	7.3 (0.54)	7.4 (0.50)	6.7 (0.50)	7.0 (0.46)	7.4 (0.46)	6.8 (0.44)	.44

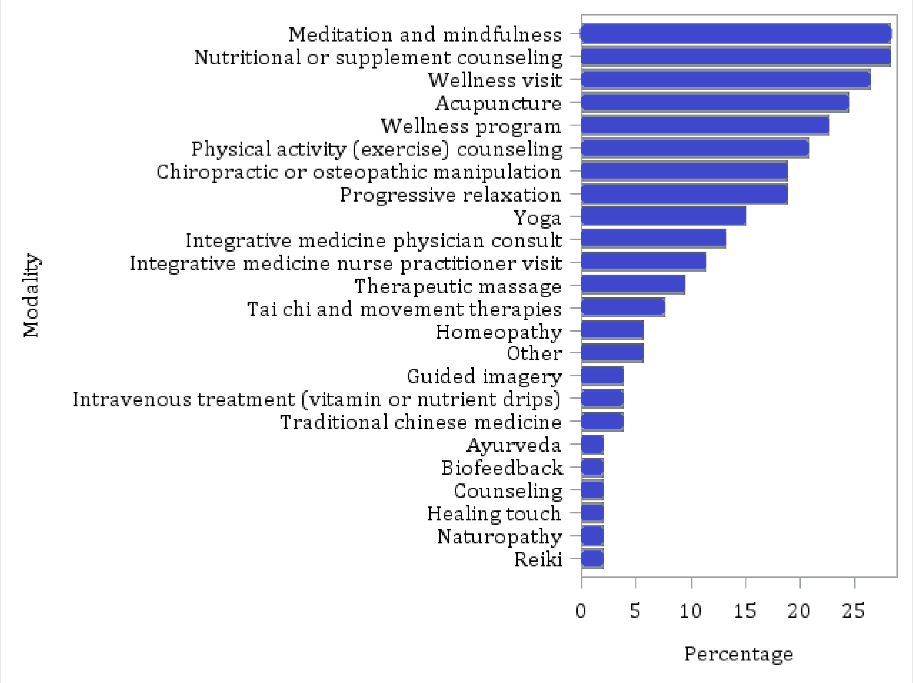
[a] PHCS-2: 2-item Perceived Health Competence Scale.

[b] PROMIS-10: 10-item Patient Reported Outcome Measurement Information System.

[c] PSS-4: 4-item Perceived Stress Scale 4.

Participants also reported their use of CIH modalities during the 12-week period by responding to the question: “Please list the whole health modalities you are currently engaged in. Please check all that apply.” Overall, 53/74 (72%) of respondents reported the use of at least one CIH modality during the 12 weeks. Some participants did not report participating in a CIH program or activity at the time of the project, which may be due to the time lag from participant pool identification and recruitment of individual participants. The reported use of CIH modalities increased over time, with 33/42 (79%) of respondents reporting using at least one modality in week 12, but only 7/18 (39%) of respondents reported using any modality in week 1. One participant reported using 12 different modalities in week 12. Over the 12 weeks, the majority, 33/53 (62%), of those that reported any modality, reported the use of two or more unique modalities (not including reporting of the same modality over subsequent weeks). Figure 1 shows the percentage of unique modality use reported over the 12 weeks (removing duplicate reports for the same modality by individual respondents) among those that reported any modality use. Meditation and mindfulness, nutritional or supplement counseling, wellness visit, acupuncture, wellness program, physical activity (exercise) counseling, progressive relaxation, chiropractic or osteopathic manipulation, yoga, and integrative medicine physician or nurse practitioner visit were the most frequently reported CIH modalities.

Figure 1. Frequency of reported unique modality use among those who reported use of at least one modality (n=53). The x-axis represents the percentage of patients reporting the modality shown on the y-axis.



Interviews

Interview data indicated that 4 main themes emerged from our discussions with the PCPs involved in the ePRO implementation: (1) SM can be burdensome for providers (acceptability); (2) PCPs delegate SM duties to their staff, such as registered nurses and licensed practical nurses, to make this process more feasible; (3) staff within the primary care clinic are more appropriate for being involved in the ePRO survey implementation; (4) veteran patients are often challenged with using both SM and the ePRO survey, therefore making this implementation less feasible for them. Table 3 highlights representative quotes from the 5 PCPs mapped to each of these themes.

Table 3. Representative quotes from primary care providers.

Themes	Representative quotes
Secure messaging can be burdensome on providers (lack of acceptability)	• "For 6 months [I] didn't have RN so duty fell to me. Crazy and not pleasant. This past year I had a segment of time where RN was not effective so everything came to me anyway. My patients needed a response so I did it." (physician) • "I don't see it being me. I'm a float and don't have a panel…We're busy so I don't see another 16 secure messages working. It wouldn't go well for us." (physician) • "I don't mind once in a while but don't want to do it from now until I retire…I have no admin time. If there's any extra admin then I have other things and I won't be able to do it." (physician) • "If you're talking about any provider in the clinic doing that right now primary care physicians are completely and totally over the top on what we have to do. Anything you propose as an addition will not be met well." (physician assistant)
PCPs[a] delegate secure messaging duties to their staff (RNs[b] and LPNs[c]; feasibility)	• "I've had nationally as designated an RN tasked do this [help out with secure messages] so all things are filtered through there." (physician) • "I have my RN and LPN who looks at secure messages for me. LPN takes care of sending out surveys and stuff. She manages the secure messages and she'll notify me if I need to look at secure messages. She looks for me and sends out to appropriate person. Almost like triaging." (physician) • "Depends on who on the team opens the secure messages. Different teams have different ways. On my team my LPN is very efficient and skilled. She opens messages and knows to go to front clerk or RN. If the message is too long, she comes up to me and says I emailed you and want you to respond to it. She does that well." (physician) • "A pain clinic RN forwards us secure messages, we respond, and she sends it back to patient." (physician) • "My RN gets me the message and many of the messages she answers without talking to me. She'll write back and ones that RN bumps over to me I see and attend to. And I'll take care of it from there. On the flip side, I can send secure messages to any of my patients and so can my nurse. Any one of my patients on secure messaging. I can do that as can my nurse." (physician assistant) • "If it's a test result I feel ok to write to them or if it's more complicated, I'd write back. If it's not as complicated, then RN takes care of that. She has some autonomy and many times she writes back, or she asks me could you write to the patient or could you tell me what to say." (physician assistant)
Staff (RN, LPN, and MSA[d]) would need to send out a survey on patient-reported outcomes (appropriateness)	• "Support staff. A well trained MSA could do that easily. At the end of the scheduled visit can send it out." (physician) • "I don't see it being me. I'm a float and don't have a panel. A pain clinic RN forwards us secure messages, we respond, and she sends it back to patient." (physician) • "LPN takes care of sending out surveys." (physician) • "If it's a survey question the RN or LPN or the team clerk could do that." (physician assistant)
Veterans can face technological challenges (feasibility)	• "My veterans who use secure messaging are avid users. Those who don't use it, who forget and need password; they struggle and that's the only barrier I see. Non-users won't be your friends." (physician) • "Some Veterans did not know how to do this [secure messaging] and have difficulties because some Veterans do not have computers at home. Veterans don't know how to do this." (physician) • "I have quite a few elderly patients who can't use computers. Smart TV and YouTube are ok but computer is not. Certain populations also have poverty and they don't have access to a smart phone and don't want to go to the library for a computer." (physician)

[a] PCPs: primary care providers.

[b] RNs: registered nurses.

[c] LPNs: licensed practical nurses.

[d] MSA: medical support assistant

Discussion

Principal Findings

Using electronic tools for survey administration and dissemination, such as web-based survey platforms, secure email, and text messaging, has the potential to enhance the collection of ePROs by increasing the efficiency of reaching larger populations and collecting data at multiple time points without having to redirect valuable clinical time for data collection. We sought to determine the feasibility of using an SM platform tied to a web-based survey administration via Qualtrics for the collection of ePROs at 6 time points over 12 weeks within the context of VA regulations and systems. Feasibility was assessed qualitatively and quantitatively within the implementation outcomes framework using process evaluation interviews among stakeholders, the duration of time to completion, completeness of responses, and response rates. Overall, of the 100 veterans recruited to participate, 74% (74/100) of participants responded to at least one survey, with 53% (53/100) responding to more than one. However, only 1 participant responded to 5 of the 6 surveys, and only 2 responded to all 6. We learned that, whereas longitudinal ePRO data collection is possible, it is likely that participants will not respond to all surveys requested.

Weekly response rates (Table 2) increased over time, with the highest response rate obtained in week 12 at 42.0% (42/100). This is in keeping with previously published ePRO response rates [20,21,33]. There are a couple of possible explanations for this observation of later surveys achieving higher response rates. First, there was a technical error in survey links sent during the first few weeks of data collection. These technical errors included dead URLs, which decreased response rates. During weeks 6 through 8, a technical error was reported by some participants who called and said they could not open the link. Upon investigation, some users experienced dead short URL links. There were no noted patterns or similarities to participants that called to report such errors. We changed to a universal short URL–generating website. No further calls from participants reporting dead short URLs were received after changing to the short URL–generating site. Such errors clearly impacted the response rates. We learned that ensuring robust testing before the implementation of an ePRO data collection process is extremely important. Survey links were sent weekly for the first 4 weeks. Participant confusion about the completion of multiple surveys and which surveys still needed to be completed may have negatively impacted response rates, particularly when there was short spacing between survey requests. Although the majority of completed surveys were received within 1 week of the SM request, late surveys may have overlapped with subsequent SM requests. The increased response rate in week 12 may, in part, be due to the greater elapsed time between requests.

We found that most surveys (150/175, 85.7%) were complete for all items from the 3 scales, and of those missing items, the majority (17/21, 81%) were only missing 1 item. Although the degree of missing data was small, we learned that this could inform the sample size needed for future research studies. We did not observe a significant change in the computed scores from the scales over the 12 weeks (Table 2). This may be due to a wide variation in modality use among respondents, including that some may have been inconsistent or not using CIH services during the data collection. We did not have baseline measures obtained before the receipt of initial CIH services to evaluate the pre-post effect of CIH on PRO. With a convenience sample size of 100, it is not possible to analyze the effects of individual CIH interventions. Some participants engaged in multiple CIH interventions, making it difficult to isolate the effect of individual interventions. In addition, we learned that the scales used (PHCS-2, PROMIS-10, and PSS-4) may not have been best suited to capture changes in biosocial constructs and whole-person wellness. For example, in the Self-Assessment of Change questionnaire, the 18 paired terms self-assessment measure, assesses not only physical, cognitive, and emotional characteristics but also social, spiritual, and whole-person characteristics. However, the original publication on the PSS-4 [28] indicated that this was a valid measure of stress, and despite being frequently used, recent publications have suggested that the scale may lack internal consistency [34,35]. We learned that future implementation of ePRO data collection should consider utilizing other stress scales.

The small number of participants may also have contributed to the inability to detect any change over time. Only 53 participants reported data at more than one time point. Allowing survey links to *expire* would help to reduce overlap and prevent late responses. Secure messages are available in the MyHealtheVet portal, requiring participants to access the portal to see their messages. Participants who did not access the portal regularly did not receive their messages in a timely manner. In addition, the staff time required to send SM was burdensome. We learned the burden, due to the constant follow-up by the staff to ensure that messages were being received, was a major limitation. The staff had to keep track of when the last survey link was sent using a customized Microsoft Access database and manually send the next link by copying/pasting the Qualtrics auto-generated link into the SM.

It is important to recognize the unique distribution of sex in this participant group. Our group overwhelmingly comprised female veterans. The national veteran population is 10% female [36]. This bias affects the ability to generalize project findings to the broader veteran population. Women tend to be more likely to respond to web-based surveys [37,38]. Our response rate may well have been influenced by our mostly female group. Additionally, female veterans are more likely to participate in CIH activities [39].

This quality improvement project has limitations. First, using a convenience sample may not provide the representativeness that a random sample may provide. Our time from participant pool identification to recruitment was extended, potentially missing opportunities to engage participants. A timelier process for reaching out to the participant pool may have provided a more robust engagement. The personal connection between study teams and participants tends to improve responsiveness [33,38]. Participants did not have in-person contact with the study team. This may decrease responsiveness to surveys [40]. The project experienced several technical glitches (described earlier) that could have been avoided by testing processes before implementation. The project also did not collect data that may have further informed analysis (eg, frequency of computer use to assess comfort with technology use). Future projects may benefit from timely, personal outreach to potential participants, assessment of comfort with technology, and vigorous testing of processes before implementation.

Project findings and lessons learned can inform future research. Future studies should explore the means of dissemination of survey requests that have the capacity for automation to reduce the potential for human and technical errors and reduce workflow burden, which are more efficient for both the staff and participants. In addition, the system should be able to contact participants using their preferred means of communication, whether that is an email address or a text message. Future projects should also examine the optimal spacing of ePRO longitudinal data collection to optimize sensitivity to measures.

Conclusions

We demonstrated the feasibility of using SM for ePRO data collection among veterans who have received CIH services; although it is possible to use SM for ePROs, in the system's current state, it is not practical. Lack of automation, workflow burden, and potential for human error make SM a cumbersome system to use when attempting to collect repeated measures from large participant cohorts. Systems that offer customizable features to automate administration on schedule are needed to reduce the provider workflow burden and potential for human error. These results can help inform future studies, including sample size considerations, best practices for workflow and automation, and ideal characteristics for messaging systems.

Acknowledgments

The Department of Veterans Affairs, VHA, and Office of Research and Development supported the development of this study. This study was supported in part by the Complementary and Integrative Health Evaluation Center, a Quality Enhancement Research Initiative within the Department of Veterans Affairs (VA), and the Office of Patient Centered Care and Cultural Transformation, PEC 16-354. The contents of this paper do not represent the views of the Department of Veterans Affairs or the US government.

Conflicts of Interest

None declared.

References

1. Veterans Affairs. Whole Health URL: https://www.va.gov/PATIENTCENTEREDCARE/features/Expanding_the_VA_Whole_Health_System.asp [accessed 2019-04-15]
2. 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]
3. Rotondi AJ, Eack SM, Hanusa BH, Spring MB, Haas GL. Critical design elements of e-health applications for users with severe mental illness: singular focus, simple architecture, prominent contents, explicit navigation, and inclusive hyperlinks. Schizophr Bull 2015 Mar;41(2):440-448 [FREE Full text] [doi: 10.1093/schbul/sbt194] [Medline: 24375458]
4. Cohen AN, Chinman MJ, Hamilton AB, Whelan F, Young AS. Using patient-facing kiosks to support quality improvement at mental health clinics. Med Care 2013 Mar;51(3 Suppl 1):S13-S20 [FREE Full text] [doi: 10.1097/MLR.0b013e31827da859] [Medline: 23407006]
5. PRWeb. 2019. Pioneers of Whole Health Applaud Veteran's Administration Use of Whole Health and Whole Person Care as Model for Future VA Healthcare Delivery URL: https://www.prweb.com/releases/2018/01/prweb15070456.htm [accessed 2019-04-15]
6. Krejci LP, Carter K, Gaudet T. Whole health: the vision and implementation of personalized, proactive, patient-driven health care for veterans. Med Care 2014 Dec;52(12 Suppl 5):S5-S8. [doi: 10.1097/MLR.0000000000000226] [Medline: 25397823]
7. Barrett B, Marchand L, Scheder J, Plane MB, Maberry R, Appelbaum D, et al. Themes of holism, empowerment, access, and legitimacy define complementary, alternative, and integrative medicine in relation to conventional biomedicine. J Altern Complement Med 2003 Dec;9(6):937-947. [doi: 10.1089/107555303771952271] [Medline: 14736364]
8. Boon H, Verhoef M, O'Hara D, Findlay B. From parallel practice to integrative health care: a conceptual framework. BMC Health Serv Res 2004 Jul 1;4(1):15 [FREE Full text] [doi: 10.1186/1472-6963-4-15] [Medline: 15230977]
9. Nazi KM. The journey to e-health: VA healthcare network upstate New York (VISN 2). J Med Syst 2003 Feb;27(1):35-45. [doi: 10.1023/a:1021005111996] [Medline: 12617196]
10. Chumbler NR, Haggstrom D, Saleem JJ. Implementation of health information technology in veterans health administration to support transformational change: telehealth and personal health records. Med Care 2011 Dec;49(Suppl):S36-S42. [doi: 10.1097/MLR.0b013e3181d558f9] [Medline: 20421829]
11. Pinney S, Otobo E, Freeman R, Rogers J, Fasihuddin F, Ramireddy K, et al. Use of electronic patient reported outcomes and automated devices for heart failure disease management. iProc 2017 Sep 22;3(1):e24 [FREE Full text] [doi: 10.2196/iproc.8459]
12. Wintner LM, Giesinger JM, Zabernigg A, Rumpold G, Sztankay M, Oberguggenberger AS, et al. Evaluation of electronic patient-reported outcome assessment with cancer patients in the hospital and at home. BMC Med Inform Decis Mak 2015 Dec 23;15:110 [FREE Full text] [doi: 10.1186/s12911-015-0230-y] [Medline: 26699708]
13. El Miedany Y, El Gaafary M, El Aroussy N, Bahlas S, Hegazi M, Palmer D, et al. Toward electronic health recording: evaluation of electronic patient reported outcome measures (e-PROMs) system for remote monitoring of early systemic lupus patients. Clin Rheumatol 2017 Nov;36(11):2461-2469. [doi: 10.1007/s10067-017-3675-9] [Medline: 28567555]
14. Gelhorn HL, Skalicky AM, Balantac Z, Eremenco S, Cimms T, Halling K, et al. Content validity and electronic PRO (ePRO) usability of the lung cancer symptom scale-mesothelioma (LCSS-Meso) in mesothelioma patients. Support Care Cancer 2018 Jul;26(7):2229-2238. [doi: 10.1007/s00520-018-4061-0] [Medline: 29392480]

15. Ritenbaugh C, Nichter M, Nichter MA, Kelly KL, Sims CM, Bell IR, et al. Developing a patient-centered outcome measure for complementary and alternative medicine therapies I: defining content and format. BMC Complement Altern Med 2011 Dec 29;11:135 [FREE Full text] [doi: 10.1186/1472-6882-11-135] [Medline: 22206345]
16. Nundy S, Dick JJ, Goddu AP, Hogan P, Lu CY, Solomon MC, et al. Using mobile health to support the chronic care model: developing an institutional initiative. Int J Telemed Appl 2012;2012:871925 [FREE Full text] [doi: 10.1155/2012/871925] [Medline: 23304135]
17. Matthew-Maich N, Harris L, Ploeg J, Markle-Reid M, Valaitis R, Ibrahim S, et al. Designing, implementing, and evaluating mobile health technologies for managing chronic conditions in older adults: a scoping review. JMIR Mhealth Uhealth 2016 Jun 9;4(2):e29 [FREE Full text] [doi: 10.2196/mhealth.5127] [Medline: 27282195]
18. Bauer V, Goodman N, Lapin B, Cooley C, Wang E, Craig TL, et al. Text messaging to improve disease management in patients with painful diabetic peripheral neuropathy. Diabetes Educ 2018 Jun;44(3):237-248. [doi: 10.1177/0145721718767400] [Medline: 29589820]
19. Haun JN, Chavez M, Nazi K, Antinori N, Melillo C, Cotner BA, et al. Veterans' preferences for exchanging information using veterans affairs health information technologies: focus group results and modeling simulations. J Med Internet Res 2017 Oct 23;19(10):e359 [FREE Full text] [doi: 10.2196/jmir.8614] [Medline: 29061553]
20. Bae WK, Kwon J, Lee HW, Lee S, Song E, Shim H, et al. Feasibility and accessibility of electronic patient-reported outcome measures using a smartphone during routine chemotherapy: a pilot study. Support Care Cancer 2018 Nov;26(11):3721-3728. [doi: 10.1007/s00520-018-4232-z] [Medline: 29732483]
21. Wallwiener M, Heindl F, Brucker S, Taran F, Hartkopf A, Overkamp F, et al. Implementation and feasibility of electronic patient-reported outcome (ePRO) data entry in the PRAEGNANT real-time advanced and metastatic breast cancer registry. Geburtshilfe Frauenheilkd 2017 Aug;77(8):870-878 [FREE Full text] [doi: 10.1055/s-0043-116223] [Medline: 28845051]
22. Howard JS, Toonstra JL, Meade AR, Whale Conley CE, Mattacola CG. Feasibility of conducting a web-based survey of patient-reported outcomes and rehabilitation progress. Digit Health 2016;2:2055207616644844 [FREE Full text] [doi: 10.1177/2055207616644844] [Medline: 29942553]
23. Peterson K, Anderson J, Ferguson L, Mackey K. Evidence Brief: The Comparative Effectiveness of Selected Complementary and Integrative Health (CIH) Interventions for Preventing or Reducing Opioid Use in Adults with Chronic Neck, Low Back, and Large Joint Pain. Washington, DC: Department of Veterans Affairs; 2011.
24. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
25. Smith MS, Wallston KA, Smith CA. The development and validation of the perceived health competence scale. Health Educ Res 1995 Mar;10(1):51-64. [doi: 10.1093/her/10.1.51] [Medline: 10150421]
26. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, PROMIS Cooperative Group. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care 2007 May;45(5 Suppl 1):S3-11 [FREE Full text] [doi: 10.1097/01.mlr.0000258615.42478.55] [Medline: 17443116]
27. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Qual Life Res 2009 Sep;18(7):873-880 [FREE Full text] [doi: 10.1007/s11136-009-9496-9] [Medline: 19543809]
28. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
29. Bowen DJ, Kreuter M, Spring B, Cofta-Woerpel L, Linnan L, Weiner D, et al. How we design feasibility studies. Am J Prev Med 2009 May;36(5):452-457 [FREE Full text] [doi: 10.1016/j.amepre.2009.02.002] [Medline: 19362699]
30. 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]
31. Charmaz K. Constructing Grounded Theory. Thousand Oaks, CA: SAGE Publications; 2014.
32. Haun JN, Lind JD, Shimada SL, Martin TL, Gosline RM, Antinori N, et al. Evaluating user experiences of the secure messaging tool on the veterans affairs' patient portal system. J Med Internet Res 2014 Mar 6;16(3):e75 [FREE Full text] [doi: 10.2196/jmir.2976] [Medline: 24610454]
33. Sheehan K. E-mail survey response rates: a review. J Comput-Media Commun 2001;6(2). [doi: 10.1111/j.1083-6101.2001.tb00117.x]
34. Ingram PB, Clarke E, Lichtenberg JW. Confirmatory factor analysis of the perceived stress scale-4 in a community sample. Stress Health 2016 Apr;32(2):173-176. [doi: 10.1002/smi.2592] [Medline: 24995556]
35. Ezzati A, Jiang J, Katz MJ, Sliwinski MJ, Zimmerman ME, Lipton RB. Validation of the perceived stress scale in a community sample of older adults. Int J Geriatr Psychiatry 2014 Jun;29(6):645-652 [FREE Full text] [doi: 10.1002/gps.4049] [Medline: 24302253]
36. Veterans Affairs. 2019. National Center for Veterans Analysis and Statistics URL: https://www.va.gov/vetdata/Veteran_Population.asp [accessed 2019-09-27]

37. Smith G. ERIC - Education Resources Information Center. 2008. Does Gender Influence Online Survey Participation? A Record-Linkage Analysis of University Faculty Online Survey Response Behavior URL: https://eric.ed.gov/?id=ED501717 [accessed 2020-06-10]
38. Aerny-Perreten N, Domínguez-Berjón MF, Esteban-Vasallo MD, García-Riolobos C. Participation and factors associated with late or non-response to an online survey in primary care. J Eval Clin Pract 2015 Aug;21(4):688-693. [doi: 10.1111/jep.12367] [Medline: 25929295]
39. Evans EA, Herman PM, Washington DL, Lorenz KA, Yuan A, Upchurch DM, et al. Gender differences in use of complementary and integrative health by US Military veterans with chronic musculoskeletal pain. Womens Health Issues 2018;28(5):379-386 [FREE Full text] [doi: 10.1016/j.whi.2018.07.003] [Medline: 30174254]
40. Shih T, Fan X. Comparing response rates in e-mail and paper surveys: a meta-analysis. Educ Res Rev 2009 Jan;4(1):26-40. [doi: 10.1016/j.edurev.2008.01.003]

Abbreviations

CIH: complementary and integrative health
ePRO: electronic patient-reported outcome
HIT: health information technology
PHCS-2: 2-item Perceived Health Competence Scale
PRO: patient reported outcomes
PROMIS-10: 10-item Patient Reported Outcome Measurement Information System
PSS-4: 4-item Perceived Stress Scale 4
SM: secure messaging
VA: Veteran Administration
VHA: Veterans Health Administration

Edited by G Eysenbach; submitted 23.07.19; peer-reviewed by L Garvin, V Subbian, J Kahn, B Smith, J Edwards; comments to author 19.08.19; revised version received 06.12.19; accepted 27.04.20; published 26.06.20.

Please cite as:
Haun JN, Alman AC, Melillo C, Standifer M, McMahon-Grenz J, Shin M, Lapcevic WA, Patel N, Elwy AR
Using Electronic Data Collection Platforms to Assess Complementary and Integrative Health Patient-Reported Outcomes: Feasibility Project
JMIR Med Inform 2020;8(6):e15609
URL: http://medinform.jmir.org/2020/6/e15609/
doi:10.2196/15609
PMID:32589163

Original Paper

Applied Practice and Possible Leverage Points for Information Technology Support for Patient Screening in Clinical Trials: Qualitative Study

Linda Becker[1], Dipl-Psych, Dipl-Phys, Dr; Thomas Ganslandt[2,3], Dr med; Hans-Ulrich Prokosch[4], Dr rer biol hum, PhD; Axel Newe[4], Dipl-Ing (FH), Dr

[1]Chair of Health Psychology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany
[2]Department of Biomedical Informatics, Heinrich-Lanz-Zentrum, Mannheim, Germany
[3]University Medicine, Ruprecht-Karls University Heidelberg, Heidelberg, Germany
[4]Chair of Medical Informatics, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Axel Newe, Dipl-Ing (FH), Dr
Chair of Medical Informatics
Friedrich-Alexander University Erlangen-Nürnberg
Wetterkreuz 13
Erlangen, 91058
Germany
Phone: 49 91318526720
Email: axel.newe@fau.de

Abstract

Background: Clinical trials are one of the most challenging and meaningful designs in medical research. One essential step before starting a clinical trial is screening, that is, to identify patients who fulfill the inclusion criteria and do not fulfill the exclusion criteria. The screening step for clinical trials might be supported by modern information technology (IT).

Objective: This explorative study aimed (1) to obtain insights into which tools for feasibility estimations and patient screening are actually used in clinical routine and (2) to determine which method and type of IT support could benefit clinical staff.

Methods: Semistandardized interviews were conducted in 5 wards (cardiology, gynecology, gastroenterology, nephrology, and palliative care) in a German university hospital. Of the 5 interviewees, 4 were directly involved in patient screening. Three of them were clinicians, 1 was a study nurse, and 1 was a research assistant.

Results: The existing state of study feasibility estimation and the screening procedure were dominated by human communication and estimations from memory, although there were many possibilities for IT support. Success mostly depended on the experience and personal motivation of the clinical staff. Electronic support has been used but with little importance so far. Searches in ward-specific patient registers (databases) and searches in clinical information systems were reported. Furthermore, free-text searches in medical reports were mentioned. For potential future applications, a preference for either proactive or passive systems was not expressed. Most of the interviewees saw the potential for the improvement of the actual systems, but they were also largely satisfied with the outcomes of the current approach. Most of the interviewees were interested in learning more about the various ways in which IT could support and relieve them in their clinical routine.

Conclusions: Overall, IT support currently plays a minor role in the screening step for clinical trials. The lack of IT usage and the estimations made from memory reported by all the participants might constrain cognitive resources, which might distract from clinical routine. We conclude that electronic support for the screening step for clinical trials is still a challenge and that education of the staff about the possibilities for electronic support in clinical trials is necessary.

(JMIR Med Inform 2020;8(6):e15749) doi:10.2196/15749

KEYWORDS

clinical trial; patient screening; electronic support; clinical information systems; inclusion criteria; exclusion criteria; feasibility studies; mobile phone

Introduction

Background

Clinical trials are one of the most challenging and most meaningful designs in medical research. Many instructions on how to design a clinical trial optimally can be found in the literature [1,2]. However, the most challenging and essential steps before a clinical trial can start are the phases of feasibility estimations and patient prescreening. The latter includes the identification of patients who fit the study design and who fulfill the inclusion criteria but do not fulfill the exclusion criteria. The screening step (ie, both feasibility screening and patient prescreening) for clinical trials might be supported by modern information technology (IT). Hospitals and other research organizations are challenged with regard to establishing appropriate IT architecture [3].

Challenges in Patient Screening and Recruitment

It is well known that patient recruitment is a crucial factor for the success of a clinical trial and that failing to achieve recruitment objectives is a common problem [4,5]. This was first reported in 1984 [6], but it has not changed significantly until very recently [7]. For example, McDonald et al [8] found that only 31% of systematically reviewed trials recruited to 100% of their original target and that 45% failed to recruit to within 80% of the target.

A lot of research has been conducted on unveiling and discussing typical problems with regard to achieving planned recruitment goals [9]. Several strategies to improve recruitment have been analyzed [8,10-12], but the specific solutions from individual trials are not easily generalizable [5]. Some major problems are the work overload of staff and their lack of time with regard to patient recruitment [4,13]. Therefore, if electronic support (ie, support by IT, see section Prior Work on the Analysis of Workflows for Screening and Recruitment) accelerated the recruitment procedure, it could relieve the staff of some of these duties. However, a common problem is that the staff have no access to and no awareness of relevant trial information, which is often available via paper-based documents only [4,14]. A further problem is that clearly defined, unambiguous inclusion and exclusion criteria for eligibility are often missing [4]. Furthermore, inclusion and exclusion criteria are often described in free text, making the mapping to electronically available data difficult [4,14].

Prior Work on the Analysis of Workflows for Screening and Recruitment

Although many papers regarding the support of clinical trials by means of IT have been published, most of them present prototypes or stand-alone solutions [15-17]. Many authors agree that the impact on the routine workflow of the involved staff needs to be kept as low as possible and that understanding these workflows is crucial [4,18,19]. However, little research has been conducted on these routine workflows outside the setting of site-specific solutions. Furthermore, only a few studies dwell on the involved actors or roles. A precise description of a routine workflow was published by Embi et al [20,21], and it included the involved physicians, main investigators, and clinical research assistants. Another analysis of the trial management workflow of oncological phase III and phase IV trials at 2 sites in São Paulo and Rio de Janeiro was conducted by de Carvalho et al [22]. They found a lack of standardized processes for data capture, a multiplicity of data repositories, and a shortage of decision support systems. They concluded that workflows need to be reorganized to use IT more efficiently and that standard procedures need to be established. Trinczek et al [23] modeled workflows in a more formal way using the Business Process Model and Notation based on unstructured interviews and concluded that complexity could lead to redundant work by an investigator and a study nurse performing the same steps twice with the same patient.

Moreover, a strong focus on both the role of physicians and on the medical field of oncology can be observed in previous research [24]. Although the role of physicians is very important [20], they are not the only party involved in the bedside care of patients. Nurses and dedicated trial personnel also need to be considered.

Motivation and Aims

There are well known and still-existing problems with regard to screening and recruiting patients for clinical trials. Clinical information systems (CIS), electronic health records, and, more generally, hospital information systems (HIS) are often considered as suitable tools for supporting the process of trial management [4,9,18]. Any IT-based solution must fit into the respective workflows [4,18,19], but these workflows—and in particular, the routine of the bedside staff that use these IT systems—have largely remained without investigation so far.

This qualitative and explorative study aimed (1) to evaluate which tools for feasibility estimations and patient screening are actually used in clinical routine and (2) to evaluate which method and type of IT could support the clinical staff. The findings are intended to lay the foundation for a larger, more representative, and more structured study.

We focused on the real-world implementation of workflows, regardless of theoretical or predefined models, as described in the study by Lee et al [14]. To achieve these goals, we conducted semistandardized interviews in a German university hospital.

Keeping in mind that trial management is a complex process and likely to require individual approaches fitted to the respective setting, we aimed to include a variety of medical disciplines and a variety of medical staff. Therefore, we considered personnel who were directly involved in screening and recruiting patients for clinical trials, with a strong focus on the bedside staff. We assumed that the bedside staff are usually burdened with nontrial-related duties and, therefore, are most likely hampered by suboptimal workflows.

Methods

The report of this study is based on the Consolidated Criteria for Reporting Qualitative Research [25].

Recruitment of Interviewees

To achieve the best possible variation (in age, gender, role, experience, and medical discipline), all clinics of a German university hospital were contacted by email, the project was briefly outlined, and participation was kindly requested. Of the 25 clinics, 11 responded, and 5 out of these 11 declined (either because of studies not being undertaken at the facility or without further explanation). Appointments were arranged with 5 of the remaining 6 clinics; 1 clinic canceled later.

Interviews were conducted with 5 participants (1 per clinic, 2 males, mean age 37.2, SD 7.9 years), each from a different ward (cardiology, gynecology, gastroenterology, nephrology, and palliative care). One of the interviewees was personally known to author AN on a professional level from a previous collaboration. Out of the 5 participants, 4 were directly involved in patient screening. Three of the interviewees were physicians, 1 was a research assistant, and 1 was a study nurse. The variety of interviewees was deemed sufficient for a qualitative study; therefore, no further efforts were made to increase the number.

Written and informed consent was obtained from all participants. The local ethics committee of Friedrich-Alexander University Erlangen-Nürnberg approved the study.

Data Collection

Semistandardized interviews that comprised 4 parts were conducted in a one-to-one setting (interviewer-interviewee). The location was chosen by the interviewees; in all cases, it was an undisturbed office environment. All participants had taken sufficient time and were not under time pressure. Each interview lasted approximately 30 min and was conducted in German by author AN, who was a PhD student in medical informatics at the time of the interviews. All interviewees were native German speakers. Before the interviews, the participants were informed about the purpose of the interview and the research context.

The interviews were voice recorded electronically using the built-in recording app of a smartphone and were then transcribed and anonymized before analysis. Written notes were not taken, but a short questionnaire with demographic variables was filled out by the interviewees. The interviews were divided into the following 4 parts: (1) the actual state of study feasibility assessment, (2) the actual patient prescreening strategy, (3) the actual IT support for feasibility estimation and prescreening, and (4) the request for IT support. The interview guidelines can be found in Multimedia Appendix 1. According to the qualitative approach of this study, no a priori hypotheses were developed or considered subsequently.

Data Analysis

The interview data were analyzed by 2 independent raters (authors AN and LB)—native German speakers— who were both postdoctoral researchers (author AN, male: medical informatics and author LB, female: health psychology) at the time of the analysis.

The analysis was based on the original German transcripts. It was carried out as follows. First, both raters independently identified and coded statements (quotes) in the transcripts that belonged to one of the research questions. Second, the coded statements were compared, and matches among the interviewers were collected in a separate document. Mismatches were discussed until consensus was reached. Next, based on discussions between the authors, clusters of similar statements were determined, and general terms for the categories were agreed upon. Subsequently, the frequencies of each category were determined. These categories as well as exemplary statements are provided in the Results section.

Finally, German quotes were translated into English for publication. (Note: Naturally spoken language is difficult to translate, especially if it comes from free speech. Therefore, the translated quotes may be *bad English*. However, they were already *bad German* in the first place, and the authors intended to keep the authenticity of the quotes, which naturally goes hand in hand with linguistic errors.) The data analysis, however, was based on the original German texts and was performed by native German speakers only.

Results

The original interviews were transcribed to 27,985 words, from which 193 key statements (quotes) were extracted. All citations and classifications are provided in Multimedia Appendix 2.

User Statistics

Out of the 5 interviewees, 4 were directly involved in patient screening. The other interviewee was a clinician who was responsible for study co-ordination and who delegated the screening to a study nurse. The percentage of work time that was spent screening for study participants ranged between 1% and 100% (mean 28.7%, SD 40.4%). The number of actual clinical trials ranged between 0 for the research assistant and 25 for 1 clinician (mean 12.8, SD 11.0). The number of inclusion and exclusion criteria per study ranged between 10 and 28 for each. The percentage of standard inclusion and exclusion criteria that remained the same for each study was estimated to be approximately 70% (range 50%-100%). The average number of patients who had to be recruited within 1 month was estimated to be 30.8 (SD 31.0; minimum 5, maximum 73, and 1 missing). A high-level coordination office was present in 3 of the 5 wards.

First Interview Part: Study Feasibility—Current Situation

In the first part of the interviews, the participants reported how they get to study feasibility estimations, that is, how they assess if there are enough patients available who fulfill the inclusion criteria but do not fulfill the exclusion criteria. An overview of the categorized answers and their frequencies is provided in Table 1.

Table 1. Answer categories in part 1 of the interviews in which the interviewees reported how they get to study feasibility estimations (N=5).

Category number	Category	Value, n[a] (%)
1.1	Experience	3 (60)
1.2	Internal statistics	3 (60)
1.3	Gut feeling	3 (60)
1.4	Works (very) well	3 (60)
1.5	Estimations	2 (40)
1.6	From memory, in mind	2 (40)
1.7	Ask colleagues or other wards	2 (40)
1.8	Literature search	2 (40)
1.9	Automatically from memory	1 (20)
1.10	Searching in protocols	1 (20)
1.11	Extrapolation from previous years’ data (problem: bad documentation so far)	1 (20)
1.12	Parallel to the clinical routine	1 (20)
1.13	Personal exchange between clinicians, coordinators, central coordinators, and bedside staff	1 (20)
1.14	Search in databases if the study is important	1 (20)
1.15	Very time consuming	1 (20)
1.16	We have no search engine	1 (20)
1.17	Looking in existing pool of patients	1 (20)
1.18	Difficulties	1 (20)
1.19	Underestimations	1 (20)
1.20	Sometimes overestimations and sometimes underestimations	1 (20)
1.21	Sometimes good, sometimes bad, or sometimes average	1 (20)
1.22	Need for exact estimates	1 (20)
1.23	Problems when not documented	1 (20)
1.24	No quality management	1 (20)
1.25	Error prone	1 (20)
1.26	Pessimistic guessing	1 (20)
1.27	Create a documentation of included and excluded patients	1 (20)

[a] The frequencies indicate the number of interviewees out of 5 who gave answers that fit into the category.

All interviewees reported that most of the feasibility estimation is done from memory or is a *gut feeling* and that it is mostly based on experience:

> *Everyone goes through the complete patient lists in his mind, you sometimes have redundancies, but nevertheless you get a very good result.* [Quote #1]

Furthermore, it was reported that a comparison with previous studies is performed from memory or in databases:

> *I just look at the numbers of [year] and do my queries, using Access.* [Quote #24]

This offers good results for studies with similar inclusion and exclusion criteria, but it works rather poorly for new types of investigations with different criteria. Two of the interviewees rated the actual procedure as good.

Only 1 interviewee reported an ongoing documentation of feasibility estimations in a separate file:

> *Then I would create a separate documentation in research, where I document all the patients that we have on the ward and I just for each patient then note to what extent the inclusion and exclusion criteria applied that I have come to the conclusion that we are trying to integrate or not so that later.* [Quote #35]

Furthermore, the study nurse reported that screening lists are created in which all screened patients and the reasons for inclusion or exclusion are listed. One interviewee reported that the current procedure tends to result in underestimations. Another interviewee stated that it depends on the study of how good the procedure works and if it results in underestimations or overestimations. The other interviewees did not comment on this.

Most of the participants reported an active exchange among study nurses and clinicians from the same ward and from other wards for assessing feasibility estimations irrespective of the responsibilities (ie, the study nurses reported that they ask the clinicians and vice versa). Moreover, 1 interviewee reported extrapolation from previous years' data based on specific databases. However, he also mentioned that some therapy situations have not been well documented, which makes the feasibility estimations difficult.

Furthermore, 1 interviewee reported doing an extensive literature search for getting prevalence rates and extrapolating this for the patient numbers in the actual department:

> *Then we extrapolate the current year.* [Quote #11]

Overall, the actual state of study feasibility estimation was dominated by human communication and estimations from memory. Success mostly depended on experience. Electronic support had only little importance so far.

Second Interview Part: Patient Screening—Current Situation

In the second part of the interviews, participants were asked how eligible patients for recruitment were identified for active clinical trials. An overview of the answer categories and the frequencies is provided in Table 2.

Again, most of the participants reported that most of the screening procedure is done from memory:

> *Because it all happens in mind.* [Quote #49]

And that it is best to memorize every study as well as all inclusion and exclusion criteria:

> *It is best if I know all the studies that are currently running in our ward.* [Quote #77]

Moreover, 1 interviewee stated:

> *In principle, it all depends on me, both the selection, the thinking about which patient exists, which patient might be suitable, contacting patients, including patients yes or no, and continuing to look after the patients, all my job.* [Quote #53]

Three interviewees felt that screening of patients requires much effort, as expressed by 1 of them:

> *That is an additional effort, exactly, which is usually not paid for. [...] So, it is usually the case that these studies run alongside the normal work of the doctors. They are all working to capacity anyway.* [Quote #61]

One interviewee stated that for restrictive inclusion criteria, success is "a matter of luck" [Quote #74].

Furthermore, 2 of the interviewees named announcements in local newspapers and postings in local offices as a strategy for recruiting. This procedure is chosen, especially, if an external sponsor is involved. However, the success rate is rather low because in most cases, only a tiny fraction of the participants actually responds to newspaper announcements. However, 2 of the interviewees reported that some of the patients come on their own initiative.

Three interviewees reported that the most important and successful strategy for recruitment is (1) asking the inpatients or outpatients during regular ward rounds or (2) through personal conversation with colleagues (Quote #96: "...quite simply the personal conversation on the ward.").

In contrast, 1 interviewee reported no division of labor and that there is 1 dedicated employee per study, who is responsible for the entire workflow. He also reported that an alert from clinical personnel does not work well:

> *You have to take care of yourself every day, you get no patients reported [...].* [Quote #100]

Regarding IT usage, 1 interviewee stated:

> *Earlier it worked, and recruiting many years ago was [...] significantly better without all the systems.* [Quote #93]

Two of the interviewees reported that they actively search in the CIS. One reported searching in paper-based patient records or in Microsoft Word documents (when data are not available in the CIS and because free-text searches are not possible in the CIS) and that the applied strategy and effort depend on the study:

> *We have a wide variety of examinations, and then every single examination really matters, what kind of people do I actually need? The search for patients is correspondingly time-consuming.* [Quote #55]

One clinician reported that he uses a *study book* for some studies, which is kept up to- date by the clinicians and to which the study nurses have access to:

> *This is a study book, where the current studies from each year are always included. It is reissued once a year. There are inclusion and exclusion criteria.* [Quote #82]

Three interviewees mentioned paper-based reminder notes with the inclusion and exclusion criteria, which can be found in the treatment rooms. Another screening strategy mentioned by the study nurse was that the clinician asks the team if there is an eligible study for a specific patient. Furthermore, it was reported that regular team meetings and regional meetings take place in which inclusion and exclusion criteria are discussed and where it is decided which patients are potentially eligible for inclusion:

> *But especially for the prescreening nothing is documented, that is, everyone does it for himself [...] and thinks about how many patients should be included, who that would be, and this will then be gathered in the team meeting. During brainstorming, everybody thinks about who should be included and at a team meeting, which we have relatively often, which we always have regularly, all patients that could be included are put forward.* [Quote #50]

Table 2. Answer categories in part 2 of the interviews in which the interviewees were asked about the actual state of their screening strategy (N=5).

Category number	Category	Value, n[a] (%)
2.1	This is done in mind	4 (80)
2.2	Personal contact, actively asking ambulatory and inpatients	3 (60)
2.3	The clinician asks the team whether there is an eligible study for a specific patient	3 (60)
2.4	Personal motivation is the most important factor for successful screening	3 (60)
2.5	Printed IC[b] and EC[c] as reminder notes, handouts	3 (60)
2.6	Works well	3 (60)
2.7	Much effort	3 (60)
2.8	Regular team meetings	2 (40)
2.9	Patients come on their own	2 (40)
2.10	Active search in CIS[d]	2 (40)
2.11	Works not well	2 (40)
2.12	Announcements in local newspapers	2 (40)
2.13	Postings and flyers in local offices	1 (20)
2.14	Internally filtering of the inpatients (in mind)	1 (20)
2.15	It all depends on me	1 (20)
2.16	Search in paper-based records	1 (20)
2.17	Written documentation of patient screening strategy and reason for inclusion or exclusion	1 (20)
2.18	Initiated by sponsors	1 (20)
2.19	Not using the CIS	1 (20)
2.20	Announcements in specialist journals	1 (20)
2.21	By the sponsors themselves	1 (20)
2.22	Preselection by the study nurses	1 (20)
2.23	Not much effort	1 (20)
2.24	No regular team meetings	1 (20)
2.25	Error prone, cannot have all in mind	1 (20)
2.26	Depends on the study	1 (20)
2.27	50% in mind	1 (20)
2.28	Excel sheets with contact information	1 (20)
2.29	Matter of luck	1 (20)
2.30	Cooperation with residents	1 (20)
2.31	Scheduling program	1 (20)
2.32	Study book	1 (20)
2.33	Printing out the study book entries	1 (20)
2.34	Back then, it worked better (without IT[e])	1 (20)
2.35	Previously known patients	1 (20)
2.36	I am solely responsible	1 (20)

[a] The frequencies indicate the number of interviewees out of 5 who gave answers that fit into the category.

[b] IC: inclusion criteria.

[c] EC: exclusion criteria.

[d] CIS: clinical information systems.

[e] IT: information technology.

Only 1 interviewee reported a written documentation of the patient screening strategy, including the reason for inclusion or exclusion. Others stated that screening protocols would definitively be helpful, but they are not feasible because of time pressure. One interviewee stated that the burden for patient screening depends on the study: inpatients with high care expenses are good to be screened, but screening in clinical routine is often forgotten.

Most of the interviewees stated that the most relevant factor for successful screening is personal motivation and not the specific tools that are used for this reason:

> *Extremely important [...] it takes a lot of heart and soul.* [Quote #77]

To summarize, the actual state of the screening procedure was dominated by human communication and estimations from memory. Electronic support was used but with little importance so far. Overall, most of the interviewees saw the potential for improvement, but they were also largely satisfied with the outcome of the current approach:

> *I don't think it's very modern to do it that way.* [Quote #98]
>
> *I think that patient identification works well for us. I'm not even dissatisfied with it. I believe that this is also a cumbersome way and very time consuming, but ultimately the result fits the outcome.* [Quote #104]

Third Interview Part: Information Technology Support—Current Situation

In the third part of the interviews, interviewees were asked which kind of IT support they use in clinical routine and, especially, for patient screening. An overview of the answer categories and frequencies is provided in Table 3. All participants reported some kind of IT support. Four interviewees reported that they regularly search in the CIS. However, 1 interviewee stated that he does not use electronic systems in most cases:

> *No, not in clinical information systems.* [Quote #120]
>
> *...we're not searching in [CIS]. So, I don't think that has ever been done in our house...I've never tried that myself, and I think in our ward...nobody does that.* [Quote #122]

However, he sometimes searches in Excel (Microsoft) lists (databases) for specific diagnoses:

> *Then there is a small database for each clinical picture; partially it's just some Excel lists or something.* [Quote #121]

Furthermore, he used an electronic data capture system for 1 study, but this is not permitted for other studies because of data protection policies:

> *We only used it relatively rarely, and there are political reasons for that. [...] Because, we have assured the patient that we will not pass on their data so that we cannot simply give it to anyone [...]. So, we explicitly promised the patient that we would not do this.* [Quote #127]

Two more interviewees reported having ward-specific patient registers for specific diagnoses and with comprehensive entries (eg, blood samples) as well:

> *We have a [specific diagnosis] register, there the patients are recorded relatively comprehensively, with blood samples and everything. And then you can also research it. We all have it in there.* [Quote #105]
>
> *I just print out the whole [regular meeting] up here on a sheet of paper and go through all the patients. Most of the time I select over 60 patients.* [Quote #87]

These have been specially built for clinical trial screening by in-house research groups. The main reason why 1 interviewee does not use electronic databases regularly is that the inclusion and exclusion criteria are very complex and:

> *Could be operationalized [in principle], but [that] would be an insane effort.* [Quote #110]

The main problem that he has with the CIS is that the diagnoses do not necessarily match the diagnoses in the paper-based records:

> *I have of course already selected according to diagnoses in [CIS]. However, these diagnoses are not necessarily the diagnoses that I now have in a doctor's letter. That is the problem.* [Quote #115]

However, he stated that:

> *The good thing is...what we are always interested in...for example that the whole blood values can be seen at a glance. This is important for us if we go through such inclusion and exclusion criteria.* [Quote #119]

However, he also stated that:

> *In the end it is always the case that we have to read the doctor's letter again.* [Quote #117]

One interviewee stated that solid numbers (eg, from blood samples) make little sense in his case because operational reports are of greater importance. He also noted that he searches only sparsely in the CIS. He mentioned that there is a lot of data in free-text medical reports, but screening the whole text is too time consuming and that the doctor's letters "...cannot be evaluated at all" (Quote #111). Another problem that was mentioned is that a specific diagnosis is not documented in many cases. Moreover, it is not documented if the patient already participates in another study.

Table 3. Answer categories in part 3 of the interviews in which the interviewees reported about the actual state of information technologies support for patient screening and feasibility estimations (N=5).

Category number	Category	Value, n[a] (%)
3.1	Active search in a CIS[b]	4 (80)
3.2	Search in electronic records (Word documents), directory with findings from the examination	3 (60)
3.3	Search in databases (eg, Excel files)	3 (60)
3.4	Ward-specific patient register (very extensive)	3 (60)
3.5	Much effort	2 (40)
3.6	Not enough or not much data are collected electronically	2 (40)
3.7	Electronic patient lists	1 (20)
3.8	Beneficial	1 (20)
3.9	We do not search in the CIS	1 (20)
3.10	Do not know the CIS	1 (20)
3.11	Complex data not in the database	1 (20)
3.12	Do not have a database	1 (20)
3.13	Problem: diagnosis in the doctor’s letters does not match the entry in the CIS	1 (20)
3.14	At the end, using the (paper-based) doctor’s letters	1 (20)
3.15	Ward-specific solution	1 (20)
3.16	Concerns with data protection policies	1 (20)
3.17	Electronic scheduling program	1 (20)
3.18	In the CIS, certain information is taken over from the last entry	1 (20)
3.19	Database with recruiting numbers	1 (20)
3.20	Feasibility estimations in internal database	1 (20)
3.21	Problem: do not have access to the CIS	1 (20)
3.22	Laboratory-specific database	1 (20)
3.23	No electronical doctor’s letters	1 (20)
3.24	Milestone	1 (20)
3.25	Difficult at the beginning	1 (20)
3.26	Works well	1 (20)

[a]The frequencies indicate the number of interviewees out of 5 who gave answers that fit into the category.

[b]CIS: clinical information systems.

Furthermore, it was reported by 1 interviewee that at his ward they:

> *Produce recruitment numbers from our studies once a month, where we have our clinical database where all study patients are registered.* [Quote #134]

One problem that was raised by 1 of the interviewees was that he had difficulties getting access to the CIS and other databases:

> *We [...] have difficulty accessing this electronic data.* [Quote #138]
>
> *Often we don’t get any rights to see this, so even if we ask that we only have read rights—we don’t want to document anything in the patient record, that’s totally okay, but we would like to be able to read it.* [Quote #139]

To summarize, there are already a few, mostly self-developed, solutions, but most of them are only partially used or have the potential to be improved:

> *Everything has grown historically, and these working groups have been on the road for many years and they almost always work with their own databases.* [Quote #125]

Fourth Interview Part: Information Technology Support—Request From Staff

In the last part of the interviews, the interviewees had the opportunity to express their requests for future IT support. An overview of the interview answers is provided in Table 4. The initial, spontaneous answers of 3 of the 5 interviewees were that it is not easy or not realistic to use or to develop an IT tool that is really helpful:

> *Not easy...Not easy...* [Quote #146]
>
> *I think that the things that you really need cannot be implemented at all, so they really cannot be implemented.* [Quote #169]
>
> *I don't know of any tool that would make work easier, right?* [Quote #175]

However, after the interviewees were given time to think about the potential of IT support, the interest and need for electronic support became more apparent to them. After a period of consideration, most participants stated that it would be helpful to have an electronic database where one could search and enter criteria and which creates a list with patient proposals:

> *Definitely, because we have so many patients. And if there was such a thing or if I could wish for something: we take over certain criteria that are stored operationalized.* [Quote #147]
>
> *[...] Every patient who has a main diagnosis like this, of course, has to appear somewhere in a database field and with a YES / NO query or whatever, that you can select that. But there are possibilities that would help us extremely.* [Quote #154]
>
> *For some things, I don't find it wrong to search in the hospital information system.* [Quote #171]
>
> *So, there would be many options. A clever mind would have to sit behind it and go through it individually with the colleagues from the [department] and then quasi operationalize it. And then a database. Then you could do a lightning search.* [Quote #156]

This would be appreciated for both study feasibility estimation as well as for patient prescreening.

Regarding whether self-paced or proactive approaches would be preferred, the interviewees stated that both approaches have advantages and that this depends on the study and the number of eligible patients. An interviewee said of a self-paced approach:

> *If of course you had a huge file now if I had a huge file where I could enter certain things and it would at the end vomit the patients who met all of these criteria, that would of course be fantastic.* [Quote #163]

Another interviewee said of a proactive approach:

> *It would of course be really convenient if I didn't have to search through this electronic documentation myself anymore, but if there was a programmed tool that simply queries certain fields and data at defined times or in defined periods, and whenever there is just a certain result comes out, I get a "pling" on the screen, [...] That would be extremely great.* [Quote #183]

The main concerns were too many alerts for the proactive solutions, but with a flexible alert time that depends on the study, it would be an acceptable solution:

> *Because I think the problem of these 10-minute memories or something becomes relevant if you recruited at many stations, [...] It might just be nicer if you bundled it up and said you get 2 in the morning and at noon sometimes a report bundled with everyone, but for [...] critical patients it would be totally okay for [...].* [Quote #192]

However, for patients who are time constrained it would be absolutely acceptable to get several alerts a day:

> *But if that were continuously and I was notified at all times, that would be a major advantage.* [Quote #189]
>
> *That would be very profitable, precisely because we expect fast progress and rapid changes in the general condition.* [Quote #190]

However, it was also stated that an IT system would be too time consuming and would lead to additional work:

> *Very few things will run so automatically that you make a request and that the system is completely there in its perfection, i.e. it takes a lot of time and money.* [Quote #175]

For example, it would be very important to keep the data (eg, the main diagnosis) up to date, but that does not always happen or takes too long:

> *The main diagnosis must always be kept up-to-date, and that would of course be important to keep the doctor's letters up-to-date.* [Quote #152]

Furthermore, all data would have to become operationalized, which is also not always possible:

And these are recurring inclusion/exclusion criteria in many studies, which are very similar, of which itcould be operationalized, but would be an insane effort. [Quote #110]

> *[...] There is always a certain limitation of course there is always.* [Quote #168]

A further concern was to get interdepartmental data access and not only ward-specific access:

> *[...] There would be a significant added value if we could ensure that there were cross-departmental collaborations.* [Quote #187]

Moreover, 1 interviewee complained that he does not have access to patient data from local offices with which they co-operate, and this is because of data protection policies:

> *So especially if you work together with private practices, we cannot access their data, because their computers with patient data are not connected to the network for data protection reasons.* [Quote #170]

Table 4. Answer categories in part 4 of the interviews in which the interviewees were asked for their request for information technologies support (N=5).

Category number	Category	Value, n[a] (%)
4.1	Database (tool in which some criteria [eg, main diagnosis] could be entered and which creates a list with patient proposals)	4 (80)
4.2	Proactive system	4 (80)
4.3	Passive system	4 (80)
4.4	Not easy, not realistic	3 (60)
4.5	Would be helpful/beneficial/fantastic	3 (60)
4.6	Interesting project, would like to learn more	2 (40)
4.7	I have concerns with regards data protection	2 (40)
4.8	Additional work, much time effort	2 (40)
4.9	Problem: difference between easy and complex cases/studies (number of IC[b] and EC[c]), need for specific solutions	2 (40)
4.10	Active or passive depends on the study and the number of available patients	1 (20)
4.11	Would be time saving	1 (20)
4.12	Access to data from local offices	1 (20)
4.13	Databases have to be kept up-to-date, and this is not always possible in clinical routine	1 (20)
4.14	Merging (in house) interfaces	1 (20)
4.15	(Eventually) too many alerts for the proactive	1 (20)
4.16	For time-critical patients, it would be absolutely okay to get several alerts a day	1 (20)
4.17	Voice recognition software for the creation of doctor’s letters	1 (20)
4.18	Certain limitation	1 (20)
4.19	Do not know a tool that could facilitate work	1 (20)
4.20	Do not really need it	1 (20)
4.21	It takes a lot of time and money	1 (20)

[a] The frequencies indicate the number of interviewees out of 5 who gave answers that fit into the category.

[b] IC: inclusion criteria.

[c] EC: exclusion criteria.

An interesting idea proposed by 1 of the clinicians was that it would be extremely helpful to have a voice recognition software for the creation of doctoral letters.

Overall, most of the interviewees showed interest in the topic and said that they would like to learn more about the possibilities of how IT could support and relieve them in their clinical routine:

> *...which would of course make it easier if we could get to know these systems.* [Quote #181]
>
> *So, it only benefits everyone. Win-win situation, there is nothing that would be a disadvantage.* [Quote #157]
>
> *I think it’s a very interesting project.* [Quote #158]

Discussion

Summary of Principal Findings

This study aimed to obtain insights into which tools for feasibility estimations and for patient screening are actually used in clinical routine by means of a qualitative approach. Furthermore, we were interested in finding possible leverage points for using IT to support clinical staff. Overall, the actual state of study feasibility estimation and the screening procedure were dominated by human communication and estimations from memory. Electronic support was used but with little importance so far. Searches in ward-specific patient registers (databases) and searches in CIS were reported. Furthermore, free-text searches in medical reports were mentioned. Most of the interviewees saw the potential for improvement in the actual systems and were interested in learning more about the possibilities of how IT could support and relieve them in their clinical routine.

Electronic Support for Study Feasibility

All interviewees reported that most of the feasibility estimations are done from memory and that currently IT support plays only a limited role. This is surprising because if an up-to-date database of all patients from previous years with the most relevant inclusion and exclusion criteria was available, a simple, time-saving search could give good estimations [26,27].

One problem that was often raised was that in many cases, not all exclusion and inclusion criteria are known, making it challenging to receive acceptable estimations independent of the availability of IT support. On the other hand, it was mentioned that there are sometimes too many exclusion and inclusion criteria that cannot be remembered accurately. This is an easy-to-implement aspect where electronic support could begin, for example, with a defined set of data elements [28]. Most of the interviewees reported that feasibility estimations have not been well documented so far. If an electronic search was performed, this strategy could be saved, making it possible to verify these steps later when needed (eg, for publications). Therefore, we conclude that IT support for feasibility estimations appears to be beneficial to the clinical staff, but education about the possibilities of IT support is necessary.

Electronic Support for Patient Screening

The interviewees reported that IT support currently plays a limited role in patient screening. Again, they reported that most steps are done from memory. However, all interviewees agreed that more data should be electronically available, and IT support would be very helpful if there was an up-to-date database. In this context, it should be mentioned that, before our study, the university hospital had been involved in a national project on IT-supported study recruitment and that the interviewed staff was completely unaware of this.

However, concerns about the possibility of a simultaneous search for several inclusion and exclusion criteria were raised. There was a need to search for 10 to 20 inclusion and exclusion criteria per case and to obtain a list of patients who fulfill most (but not necessarily all) of them.

Therefore, it is again concluded that familiarizing the personnel with modern IT solutions seems to be necessary. For example, Bache et al [29] developed a domain-specific query language as an interface between clinicians and stored data to facilitate this task on a technical level. A properly designed software tool has proven to be an enabler for users to correctly create and execute simple feasibility queries with only a relatively small amount of training [30]. Furthermore, several cost-benefit assessments have confirmed the benefits of electronic recruitment strategies [31,32].

Electronic Support Strategies

In general, there are three strategies for patient identification and patient recruitment for clinical trials by means of HIS [18]: (1) systems that retrospectively query existing data in an HIS (Clinical Trial Recruitment Support Systems), (2) systems that monitor the occurrence of a specific event in an HIS to create some kind of alert (Clinical Trial Alert Systems), and (3) systems that require an operator to enter appropriate data to trigger an eligibility assessment.

The interviewees in our study did not prefer any of these approaches in general and clearly stated that it depends on the study (eg, the number of inclusion and exclusion criteria and if these can be operationalized easily) and, more importantly, on the patients who should be involved. For time-critical studies and rare patient groups, a proactive alert system was clearly requested. For more common patient groups with slow changes in health status, either passive systems or summarized alerts once a day or week were requested. Therefore, as has been reported previously, one of the key challenges in the design and operation of an active system is to find an operational model that has the least possible influence on the usual workflow of the target audience [18]. In addition, the threshold must be balanced carefully between a sensitivity that produces a high number of eligible patients and a specificity that avoids alert fatigue [4,15,20,21]. Several approaches and solutions have been presented in recent years [18]. Alerts can be sent out using paging systems [15,16], mobile devices [17], or emails [33,34] or can be sent directly within the HIS or the Clinical Decision Support System [13,20,21] or with a combination of both [35,36].

Personal Motivation as a Key Factor

Most of the interviewees agreed that the success of screening depends—beside the complexity of the study and the support by appropriate IT systems—on the clinical staff and on their personal motivation. Therefore, a future challenge for the development of IT support systems seems to be finding IT solutions that motivate the staff to invest greater effort in screening without hampering their everyday work.

Limitations

Our study provides numerous new insights into the work processes of the study staff directly involved in patient screening. However, our data were collected from 1 specific German university hospital, with a low sample size of 5. Therefore, the risk of bias needs to be considered, and our findings cannot be generalized to other institutions. Therefore, the next steps should be to repeat this investigation with other hospitals and within larger samples and a more structured approach. As the interviews were very time consuming and might have affected the clinical routine, a web-based survey might be a better means for this. The results of our study provide the basis for developing such a survey.

The raw data were collected in German, and the key quotes were translated into English for this publication. Naturally spoken language is difficult to translate, especially if it comes from free speech that is prone to grammatical errors and mental leaps. Therefore, parts of the content or the intention of the translated texts may have been lost in the translation. Hence, the translated quotes from the interviews that are published in this paper should be used with caution.

Outlook

The results of our study cannot be generalized, but at least the following research questions could be derived from the findings and should, therefore, be addressed by future studies in the context of patient screening and recruitment:

- How well are the staff informed about the opportunities that IT offers in their work environment? How can the staff be educated in this regard?
- Does the staff already take advantage of IT support? Is this support based on tailored solutions or standard systems?
- How does the personal motivation of the staff affect the outcome? Would IT support have an impact (negative or positive) on this motivation?
- To what extent are estimates made from memory and why?

- How does the complexity of a study (eg, number of inclusion and exclusion criteria) affect the staff's ability to estimate patient numbers correctly? Is there a threshold above which IT support makes sense? Is there a threshold below which IT support does not make sense?
- How does the diversity of IT systems in their work environment affect the staff's ability to perform tasks for patient screening and recruitment?

In this context, the excerpts from the interviews provided in Multimedia Appendix 2 as well as the distilled answer categories provided in Tables 1-4 can serve as the basis for a questionnaire design.

Another aspect that should be addressed in future research is an inversion of the initial premise that electronic support would make the daily routine *easier* because the opposite case has not yet been considered at all. Although some interviewees expressed their concern that electronic support might not be *feasible*, none of them mentioned that it could be *obstructive*. Nevertheless, as this aspect was not explicitly addressed, it should be investigated or at least considered in follow-up studies.

Conclusions

Although it seems that IT is nearly ubiquitous, our study suggests that IT support still has limited use in the screening step for clinical trials. Our main finding was that the staff is underinformed about modern IT solutions for the support of patient screening. This lack of IT usage and the resulting work-from-memory strategy might constrain cognitive resources, which might distract from clinical routine. Therefore, we conclude that it is necessary to educate the staff about the possibilities of IT support for clinical trial screening and—in addition to conducting a large-scale, more structured study based on our findings as proposed earlier—one future research option in this direction is to develop training programs that can achieve this goal.

Acknowledgments

The authors thank Martin Ross for the transcription of the interview data. This research was funded by the Innovative Medicines Initiative (IMI) project Electronic Health Records for Clinical Research (IMI grant number: 115189). HP's work was conducted within the MIRACUM consortium which is funded by the German Ministry for Education and Research (funding number FKZ01ZZ1801A). The authors acknowledge support from Deutsche Forschungsgemeinschaft (German Research Foundation) and Friedrich-Alexander University Erlangen-Nürnberg within the funding program Open Access Publishing.

Authors' Contributions

LB headed the interview design, analyzed the data, and wrote the manuscript. TG supported the interview design and contributed to the manuscript. HP supervised the study design and contributed to the manuscript. AN designed and conducted the interviews, analyzed the data, and wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guideline and structure.
[PDF File (Adobe PDF File), 135 KB - medinform_v8i6e15749_app1.pdf]

Multimedia Appendix 2
Extracted and categorized quotes from the interviews.
[DOCX File , 62 KB - medinform_v8i6e15749_app2.docx]

References

1. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989 Apr;8(4):431-440. [doi: 10.1002/sim.4780080407] [Medline: 2727467]
2. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996 Feb;17(1):1-12. [doi: 10.1016/0197-2456(95)00134-4] [Medline: 8721797]
3. Prokosch HU. Clinical information systems and translational research: increasing the efficiency of trial feasibility and patient recruitment. In: Degoulet P, Fieschi M, Ménard J, editors. E-Santé en Perspective. Paris, France: Lavoisier; May 19, 2017:147-158.
4. Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. Int J Med Inform 2011 Jun;80(6):371-388. [doi: 10.1016/j.ijmedinf.2011.02.003] [Medline: 21459664]

5. Dugas M, Amler S, Lange M, Gerss J, Breil B, Köpcke W. Estimation of patient accrual rates in clinical trials based on routine data from hospital information systems. Methods Inf Med 2009;48(3):263-266. [doi: 10.3414/ME0582] [Medline: 19387510]
6. Charlson ME, Horwitz RI. Applying results of randomised trials to clinical practice: impact of losses before randomisation. Br Med J (Clin Res Ed) 1984 Nov 10;289(6454):1281-1284 [FREE Full text] [doi: 10.1136/bmj.289.6454.1281] [Medline: 6437520]
7. Bugeja L, Low JK, McGinnes RA, Team V, Sinha S, Weller C. Barriers and enablers to patient recruitment for randomised controlled trials on treatment of chronic wounds: a systematic review. Int Wound J 2018 Dec;15(6):880-892. [doi: 10.1111/iwj.12940] [Medline: 29927054]
8. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials 2006 Apr 7;7:9 [FREE Full text] [doi: 10.1186/1745-6215-7-9] [Medline: 16603070]
9. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Methods Inf Med 2018 Jan 17;48(1):38-44. [doi: 10.3414/ME9132] [Medline: 19151882]
10. Mapstone J, Elbourne DD, Roberts IG. Strategies to improve recruitment to research studies. Cochrane Database Syst Rev 2007 Apr 18(2):MR000013. [doi: 10.1002/14651858.MR000013.pub3] [Medline: 17443634]
11. Caldwell PH, Hamilton S, Tan A, Craig JC. Strategies for increasing recruitment to randomised controlled trials: systematic review. PLoS Med 2010 Nov 9;7(11):e1000368 [FREE Full text] [doi: 10.1371/journal.pmed.1000368] [Medline: 21085696]
12. Fletcher B, Gheorghe A, Moore D, Wilson S, Damery S. Improving the recruitment activity of clinicians in randomised controlled trials: a systematic review. BMJ Open 2012;2(1):e000496 [FREE Full text] [doi: 10.1136/bmjopen-2011-000496] [Medline: 22228729]
13. Embi PJ, Jain A, Harris CM. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: a survey. BMC Med Inform Decis Mak 2008 Apr 2;8:13 [FREE Full text] [doi: 10.1186/1472-6947-8-13] [Medline: 18384682]
14. Lee Y, Jana S, Mylavarapu T, Dinakarpandian D, Owens D. MindFlow: Intelligent Workflow for Clinical Trials in Mental Healthcare. In: Proceedings of the 45th Hawaii International Conference on System Sciences. 2012 Presented at: HICSS'12; January 4-7, 2012; Maui, HI, USA. [doi: 10.1109/hicss.2012.430]
15. Butte AJ, Weinstein DA, Kohane IS. Enrolling patients into clinical trials faster using RealTime Recuiting. Proc AMIA Symp 2000:111-115 [FREE Full text] [Medline: 11079855]
16. Weiner DL, Butte AJ, Hibberd PL, Fleisher GR. Computerized recruiting for clinical trials in real time. Ann Emerg Med 2003 Feb;41(2):242-246. [doi: 10.1067/mem.2003.52] [Medline: 12548275]
17. Chow E, Zuberi M, Seto R, Hota S, Fish EN, Morra D. Using real-time alerts for clinical trials: identifying potential study subjects. Appl Clin Inform 2011;2(4):472-480 [FREE Full text] [doi: 10.4338/ACI-2011-04-CR-0026] [Medline: 23616889]
18. Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014 Jul 1;16(7):e161 [FREE Full text] [doi: 10.2196/jmir.3446] [Medline: 24985568]
19. Bates DW, Kuperman GJ, 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]
20. Embi PJ, Jain A, Clark J, Bizjack S, Hornung R, Harris CM. Effect of a clinical trial alert system on physician participation in trial recruitment. Arch Intern Med 2005 Oct 24;165(19):2272-2277 [FREE Full text] [doi: 10.1001/archinte.165.19.2272] [Medline: 16246994]
21. Embi PJ, Jain A, Clark J, Harris CM. Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc 2005:231-235 [FREE Full text] [Medline: 16779036]
22. de Carvalho EC, Batilana AP, Claudino W, Reis LF, Schmerling RA, Shah J, et al. Workflow in clinical trial sites & its association with near miss events for data quality: ethnographic, workflow & systems simulation. PLoS One 2012;7(6):e39671 [FREE Full text] [doi: 10.1371/journal.pone.0039671] [Medline: 22768105]
23. Trinczek B, Schulte B, Breil B, Dugas M. Patient recruitment workflow with and without a patient recruitment system. Stud Health Technol Inform 2013;192:1124. [doi: 10.3233/978-1-61499-289-9-1124] [Medline: 23920898]
24. Campbell MK, Snowdon C, Francis D, Elbourne D, McDonald AM, Knight R, STEPS Group. Recruitment to randomised trials: strategies for trial enrollment and participation study. The STEPS study. Health Technol Assess 2007 Nov;11(48):iii, ix-iii,105 [FREE Full text] [doi: 10.3310/hta11480] [Medline: 17999843]
25. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. Soto-Rey I, Trinczek B, Karakoyun T, Dugas M, Fritz F. Protocol feasibility workflow using an automated multi-country patient cohort system. Stud Health Technol Inform 2014;205:985-989. [doi: 10.3233/978-1-61499-432-9-985] [Medline: 25160335]
27. Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F, Work Package 7. Piloting the EHR4CR feasibility platform across Europe. Methods Inf Med 2014;53(4):264-268. [doi: 10.3414/ME13-01-0134] [Medline: 24954881]

28. Doods J, Botteri F, Dugas M, Fritz F, EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. Trials 2014 Jan 10;15:18 [FREE Full text] [doi: 10.1186/1745-6215-15-18] [Medline: 24410735]
29. Bache R, Taweel A, Miles S, Delaney BC. An eligibility criteria query language for heterogeneous data warehouses. Methods Inf Med 2015;54(1):41-44. [doi: 10.3414/ME13-02-0027] [Medline: 24985949]
30. Soto-Rey I, N'Dja A, Cunningham J, Newe A, Trinczek B, Lafitte C, et al. User satisfaction evaluation of the EHR4CR query builder: a multisite patient count cohort system. Biomed Res Int 2015;2015:801436 [FREE Full text] [doi: 10.1155/2015/801436] [Medline: 26539525]
31. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, et al. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the electronic health records for clinical eesearch (EHR4CR) European project. Contemp Clin Trials 2016 Jan;46:85-91. [doi: 10.1016/j.cct.2015.11.011] [Medline: 26600286]
32. Dupont D, Beresniak A. Assessing the financial impact of reusing electronic health records data for clinical research: results from the EHR4CR European project. J Health Med Inform 2016;7(3):235. [doi: 10.4172/2157-7420.1000235]
33. Dugas M, Lange M, Berdel WE, Müller-Tidow C. Workflow to improve patient recruitment for clinical trials within hospital information systems-a case-study. Trials 2008 Jan 11;9:2 [FREE Full text] [doi: 10.1186/1745-6215-9-2] [Medline: 18186949]
34. Trinczek B, Köpcke F, Leusch T, Majeed RW, Schreiweis B, Wenk J, et al. Design and multicentric implementation of a generic software architecture for patient recruitment systems re-using existing HIS tools and routine patient data. Appl Clin Inform 2014;5(1):264-283 [FREE Full text] [doi: 10.4338/ACI-2013-07-RA-0047] [Medline: 24734138]
35. Afrin LB, Oates JC, Boyd CK, Daniels MS. Leveraging of open EMR architecture for clinical trial accrual. AMIA Annu Symp Proc 2003:16-20 [FREE Full text] [Medline: 14728125]
36. Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, et al. A real-time screening alert improves patient recruitment efficiency. AMIA Annu Symp Proc 2011;2011:1489-1498 [FREE Full text] [Medline: 22195213]

Abbreviations

CIS: clinical information system
HIS: hospital information system
IMI: Innovative Medicines Initiative
IT: information technology

Edited by C Lovis; submitted 02.08.19; peer-reviewed by D Kohnen, J Martin-Kerry, P Hsiao, B Schreiweis, F Lanfranchi; comments to author 16.10.19; revised version received 08.03.20; accepted 28.03.20; published 16.06.20.

Please cite as:
Becker L, Ganslandt T, Prokosch HU, Newe A
Applied Practice and Possible Leverage Points for Information Technology Support for Patient Screening in Clinical Trials: Qualitative Study
JMIR Med Inform 2020;8(6):e15749
URL: http://medinform.jmir.org/2020/6/e15749/
doi:10.2196/15749
PMID:32442156

Original Paper

Factors Influencing Doctors’ Participation in the Provision of Medical Services Through Crowdsourced Health Care Information Websites: Elaboration-Likelihood Perspective Study

Yan Si[1], MA; Hong Wu[2], PhD; Qing Liu[2], MA

[1]School of Business, Wuxi Vocational College of Science and Technology, Wuxi, China

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

Corresponding Author:
Hong Wu, PhD
School of Medicine and Health Management
Tongji Medical College
Huazhong University of Science and Technology
13 Hangkong road, Qiaokou District
Wuhan
China
Phone: 86 13277942186
Email: hongwu@hust.edu.cn

Abstract

Background: Web-based crowdsourcing promotes the goals achieved effectively by gaining solutions from public groups via the internet, and it has gained extensive attention in both business and academia. As a new mode of sourcing, crowdsourcing has been proven to improve efficiency, quality, and diversity of tasks. However, little attention has been given to crowdsourcing in the health sector.

Objective: Crowdsourced health care information websites enable patients to post their questions in the question pool, which is accessible to all doctors, and the patients wait for doctors to respond to their questions. Since the sustainable development of crowdsourced health care information websites depends on the participation of the doctors, we aimed to investigate the factors influencing doctors’ participation in providing health care information in these websites from the perspective of the elaboration-likelihood model.

Methods: We collected 1524 questions with complete patient-doctor interaction processes from an online health community in China to test all the hypotheses. We divided the doctors into 2 groups based on the sequence of the answers: (1) doctor who answered the patient’s question first and (2) the doctors who answered that question after the doctor who answered first. All analyses were conducted using the ordinary least squares method.

Results: First, the ability of the doctor who first answered the health-related question was found to positively influence the participation of the following doctors who answered after the first doctor responded to the question ($\beta_{offline1}$=.177, *P*<.001; $\beta_{offline2}$=.063, *P*=.048; β_{online}=.418, *P*<.001). Second, the reward that the patient offered for the best answer showed a positive effect on doctors’ participation (β=.019, *P*<.001). Third, the question’s complexity was found to positively moderate the relationships between the ability of the first doctor who answered and the participation of the following doctors (β=.186, *P*=.05) and to mitigate the effect between the reward and the participation of the following doctors (β=–.003, *P*=.10).

Conclusions: This study has both theoretical and practical contributions. Online health community managers can build effective incentive mechanisms to encourage highly competent doctors to participate in the provision of medical services in crowdsourced health care information websites and they can increase the reward incentives for each question to increase the participation of the doctors.

(JMIR Med Inform 2020;8(6):e16704) doi:10.2196/16704

KEYWORDS

crowdsourcing; crowdsourced medical services; online health communities; doctors’ participation; elaboration-likelihood model

Introduction

Background

The imbalance between the supply and demand for medical services has caused conflicts in the patient-doctor relationship, especially because the health awareness of patients has dramatically increased in recent years [1]. With the development of online health communities in China, an increasing number of people have begun to seek web-based health information and services [2,3], and these websites have become a useful complementation [4]. However, only 6.1% of the doctors participate in online health communities to provide medical services [5]. Medical services are not easily accessible for patients in China [6], especially for patients with serious diseases and for those living in remote areas [7]. The improvement of doctors' participation in online health communities is the key to enhancing timely services and supplementary services, which will reduce the conflicts in the patient-doctor relationship and eventually improve the overall health of the country [8]. Therefore, the primary concern of the governments and health care organizations is to increase the number of doctors involved in the provision of web-based medical services.

Crowdsourcing is widely used among organizations to obtain more and better solutions for their projects by encouraging the public to perform tasks by sharing their knowledge and skills together [9,10]. It is an emerging organizational practice that has attracted much attention over the last decade, and this pattern has also emerged in the health care field [7,11,12]. Crowdsourcing is a mode of engaging a crowd of people to achieve a common goal, for example, for solving problems by sharing the problem through questionnaires and then considering the responses of all the people in the network [13-15]. In crowdsourcing, a wide range of goals can be achieved—from idea gathering to solution elaboration [16]. Crowdsourcing is also used to survey infectious diseases by capturing the symptom data that has been submitted voluntarily [17,18]. With the rapid development of the internet, an increasing number of medical question-and-answer websites have adopted the crowdsourcing mode to find better answers to solve patients' health problems, such as *Medhelp.org* in the United States and *120ask.com* in China. These crowdsourced health care information websites are widely accepted by patients [19]. This service is a type of expert-based crowdsourced medical service [20,21], which allows patients to post an "open-call" question to undefined doctors [22] with relatively low cost [16,23]. Crowdsourced health care information websites have adopted an active crowdsourcing mode, that is, the patient has an active role, wherein he/she poses a particular medical question and solicits relevant information, knowledge, opinion, and ideas from doctors [24]. By using the crowdsourced health care information websites, patients hope to describe the symptoms and receive the diagnosis and treatment of diseases and be prescribed drugs, similar to that received in common medical services. Moreover, the patients expect that doctors who play vital roles in such services will offer answers to their questions. The most apparent feature of crowdsourced health care information websites is that more than one doctor can give answers, based on their knowledge and experience, to the same question from a single patient. Therefore, by using crowdsourced health care information websites, patients can obtain more comprehensive and better suggestions.

Previous studies have investigated the motivations behind the behavior of the participating users in posting their ideas on crowdsourcing websites [25-28]. These motivations can be divided into 2 dimensions: extrinsic motivations [25-27] and intrinsic motivations [28-30]. For the extrinsic motivations, researchers have shown that financial incentives such as monetary stimulus play an important role in the users' participating behaviors [31]. Some studies have shown that the reward is the primary source of income on the crowdsourcing platforms and this reward drives users to participate in tasks [25-27]. For intrinsic motivations, some studies have proposed that the reasons for participation in crowdsourced tasks include factors such as competency, reputation, altruism, and learning, which are the critical driving forces of the participation behaviors [28-30]. The number of downloads means attention is the motivation for users to participate in YouTube [32]. However, previous studies have mainly focused on the users' participating behaviors in other products or service fields, and only little attention has been paid to the users' participating behaviors in the medical field and in empirical research from the perspective of the information system.

We employed the elaboration-likelihood model (ELM) as the theoretical base to understand how doctors process information regarding participation in the provision of medical services in crowdsourced health care information websites. The ELM originates from social psychology and argues that individuals can change their attitudes through a dual route, namely, the central route and the peripheral route [33]. In the "central route," an individual processes information such as information quality and content through careful in-depth thinking. On the contrary, in the "peripheral route," the individual makes a decision based on less cognitive thinking and simple information cues such as monetary value [34,35]. The ELM is a dual-process theory arguing that persuasion can act via the central or the peripheral route, and it is the process of the individual's attitude change as a result of being influenced by the mental effort required for the message [33,36]. The ELM also indicates that the dual routes of decision making are moderated by the potential user's motivation to elaborate on informational messages [33,36]. Since the sustainable development of crowdsourced health care information websites depends on doctors' participation, we aimed to investigate the factors influencing the doctors' participation in providing health care information on these crowdsourced websites from the elaboration-likelihood perspective. The research questions were as follows.

1. What factors affect doctors' participation in crowdsourced health care information websites?
2. How can the question's complexity moderate the central route and peripheral route?

Research Framework and Hypotheses Development

Based on the framework of ELM, this study aims to investigate the attitude of the participating doctors toward the crowdsourced health care information websites, which is persuaded by dual-process cues, namely, central cues and peripheral cues.

Central Cues

Based on the ELM framework, central cues information is a signal of project quality, which has significant positive effects on the recipient's choice [37,38]. Information quality and review quality of products or service providers are often regarded as the central cues [39-41]. The purchase behaviors of the consumers are also considered as an important signal of the product or service quality that attract other consumers to follow and make decisions [42,43]. The reputation, ability, purchase behaviors, and review behaviors of consumers can influence the decisions of the following consumers [44]. Therefore, we hypothesize that the behaviors of the doctors who answered first would be a signal to other doctors and this could influence their behaviors. Specifically, the ability of the first doctor who answered might convey a signal of the question information, that is, the question is considered worthy of answering if the reputation/ability/review rating of the doctor who answered first is high. The doctors are free to answer any question of the patients in the question pool in these crowdsourced websites. Doctors can obtain information on the questions and the former answers, including title and reputation, especially of the doctor who answered the question first. Therefore, the following doctors' participation would be influenced by information about the doctor who answered first. The ability of the doctor who answered the patient's question first is especially crucial as the quality of the doctor's answer is considered very important in an empirical model. We hypothesize that questions that are answered first by highly competent doctors will gain more attention from other doctors, and the other doctors will be driven to participate in the provision of medical services through these crowdsourced health care information websites. Thus, our first hypothesis was called the central route hypothesis and it was as follows: The ability of the doctor who answered first has a positive effect on the following doctors' participation in crowdsourced health care information websites.

Peripheral Cues

Peripheral cues are information based on less cognitive effort such as the numbers or source characteristics that rely on shortcuts [37,45]. The reward is monetary numbers, which accord with the peripheral cue. Previous studies have explored the role of reward in the crowdsourcing field [46-48] and have indicated that financial reward is the most critical motivation, as most respondents reported that they do not perform tasks for fun or to kill time [31]. Some studies have shown that money or points have a positive effect on the user's participation in online health communities [28,49]. Further, the effects of monetary incentives on other specified crowdsourcing tasks were studied [50-53]. Thus, the participation behavior of the doctors is influenced by the monetary reward, which is listed on the question information. We believe that doctors would tend to answer questions with higher expected rewards. Thus, our second hypothesis was called the peripheral cue hypothesis, which is as follows: The reward provided by the patient has a positive effect on the following doctors' participation in crowdsourced health care information websites.

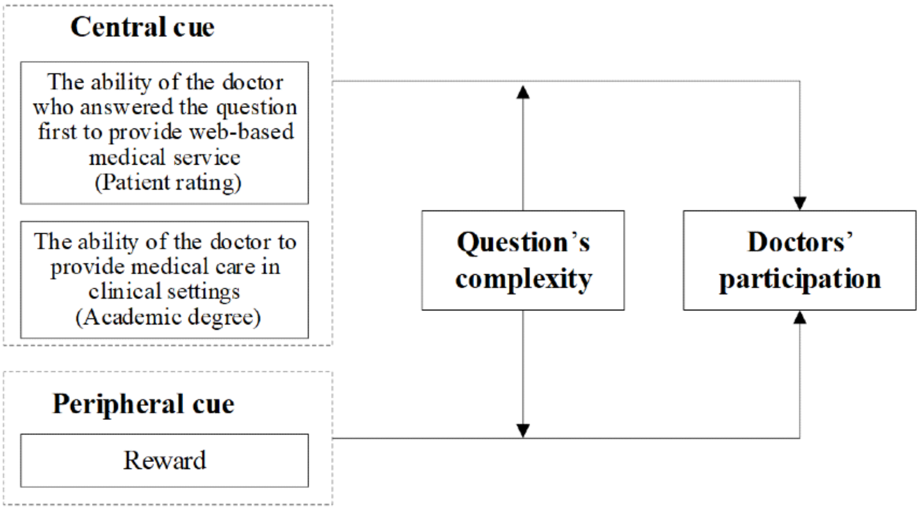
Moderating Effects

The following hypotheses are based on the moderating effects of the question's complexity on the central route and the peripheral route. The elaboration of the moderator will positively moderate the influence of the central route and negatively moderate the influence of the peripheral route [54]. Based on the ELM framework, the use of the central route and peripheral route processing for decision making is moderated by the user's ability and motivation to elaborate on informational messages [55-57], and motivation levels change the likelihood of elaboration by a user [56,57]. Patients can search/ask for information on their health problems and disease symptoms and find/ask information on the medications or other medical and health-related information in crowdsourced health care information websites. Patients with severe illness may ask more complex questions to doctors. In addition, highly complex questions can arouse the attention of doctors who have higher competencies than other doctors in clinical settings or in web-based medical services. Therefore, the solutions for highly complex questions rely more on the doctors' competency. In addition, the reward that a patient assigns is often lower than that assigned in a normal web-based health service [58]; therefore, the behavior of answering health-related questions is an act of altruism, which means that the doctor provides an answer for the "public good" [59] of the patients or health-information seekers. We believe that doctors would take the effort to solve a health problem as an act of kindness rather than for money when the question complexity is high. We hypothesized that the question's complexity has a moderating effect on the relationship between the ability of the doctor who answered first/reward and the following doctors' participation in crowdsourced health care information websites. Thus, our third hypothesis was divided into 2 categories as follows:

1. Central route processing: The question's complexity has a positive moderating effect on the relationship between the ability of the doctor who answered first and the following doctors' participation in crowdsourced health care information websites.
2. Peripheral route processing: The question's complexity has a negative moderating effect on the relationship between the reward and the following doctors' participation in crowdsourced health care information websites.

Based on the above hypotheses, the research framework is shown in Figure 1.

Figure 1. The research framework.


Central cue
The ability of the doctor who answered the question first to provide web-based medical service (Patient rating)
The ability of the doctor to provide medical care in clinical settings (Academic degree)
Peripheral cue
Reward
Question's complexity
Doctors' participation

Methods

Research Context

The 120ask website (www.120ask.com, see Figure 2) was used to obtain empirical results in this study. This website was established in 2004, and it provided a community for patients and doctors in China. The 120ask website has gathered about 5 million qualified doctors and more than 264 million patients. On this website, thousands of new health-related questions are received each day. The 120ask website is one of the top online health community platforms in China, and the monthly number of active numbers remain at above 40 million people [60]. In this platform, the doctors can share knowledge and information about the diseases and help patients improve their health conditions and receive medical diagnoses quickly and conveniently. Crowdsourced health care information websites have 2 groups of users: patients and doctors. These services allow patients to ask a health-related question to undefined doctors. The process of the crowdsourced health care information websites is as follows. First, the patient posts a question into the question pool in the online crowdsourced medical service platform with a reward and a time frame to reply. Second, within the restricted time, doctors can freely choose to answer or not and compete to win the best answer, as the reward would be given for the best answer. Third, the best answer is selected by the patient, and the corresponding doctor is granted the reward. The process of medical service provision in a crowdsourced health care information website is shown in Figure 3.

We chose the 120ask website to conduct our empirical study for the following reasons. First, the website has the history of all the consultation records saved that would help patients seek health information about similar diseases (Figure 4). Second, consultation records this website are public, which provides the patient with basic information such as gender and age. Third, the doctor's information is available on the website. Fourth, it has a large number of registered users, which enables this website to process data that are under private protection. The above features make the 120ask website a fundamentally useful website for our study.

Figure 2. The 120ask website.

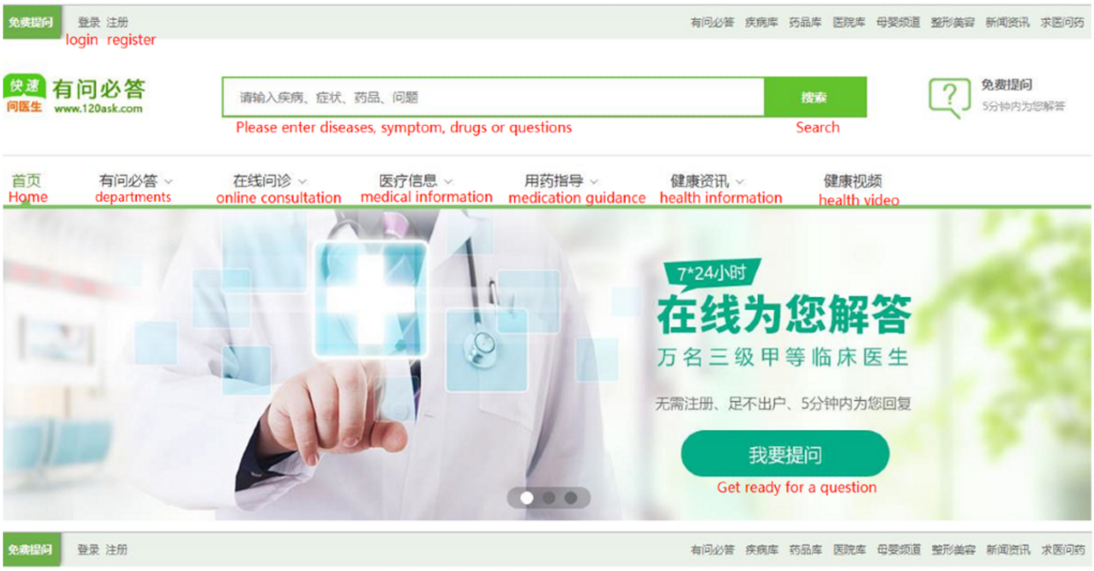


Figure 3. Process of medical service provision in a crowdsourced health care information website.

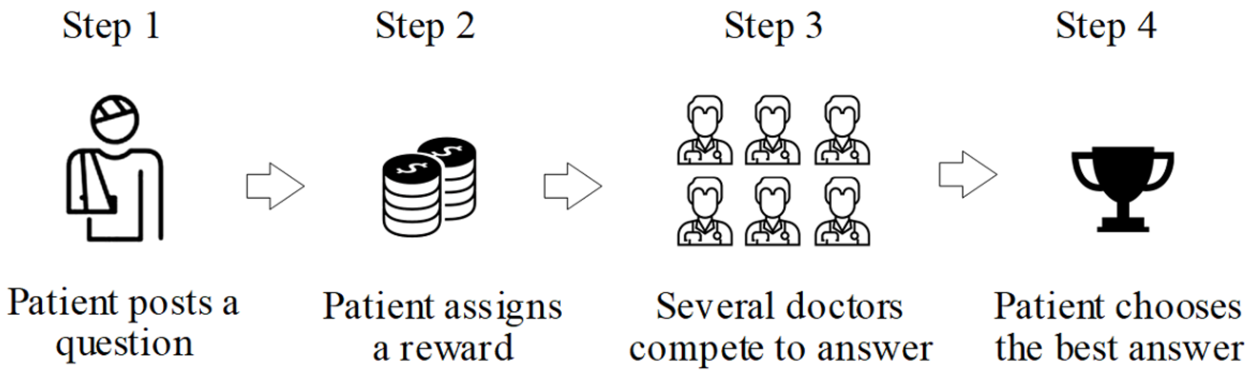
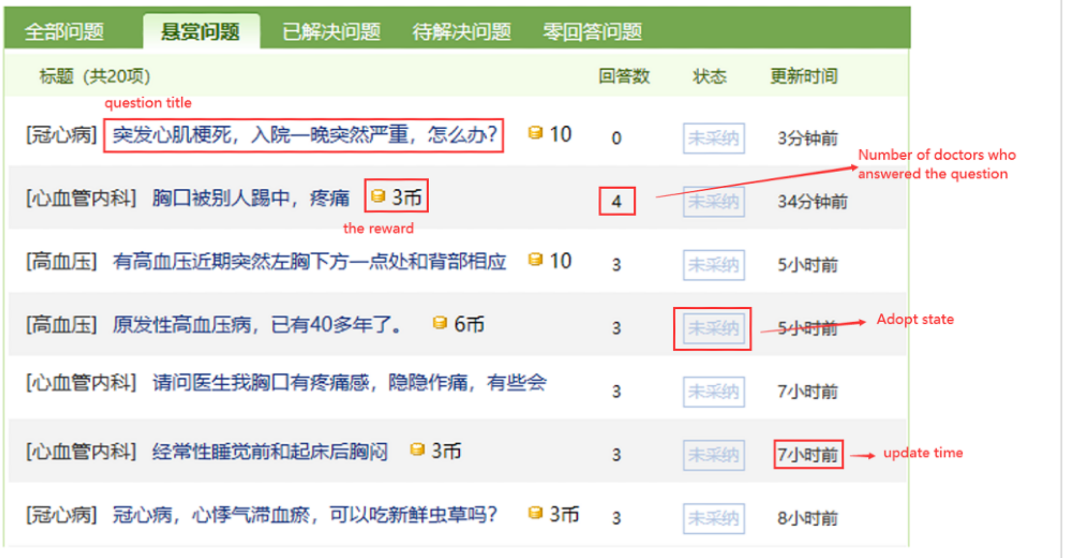


Figure 4. Parts of the history of the records in the crowdsourced health care information website.



标题（共20项）		回答数	状态	更新时间
[冠心病] 突发心肌梗死，入院一晚突然严重，怎么办?	10	0	未采纳	3分钟前
[心血管内科] 胸口被别人踢中，疼痛	3币	4	未采纳	34分钟前
[高血压] 有高血压近期突然左胸下方一点处和背部相应	10	3	未采纳	5小时前
[高血压] 原发性高血压病，已有40多年了。	6币	3	未采纳	5小时前
[心血管内科] 请问医生我胸口有疼痛感，隐隐作痛，有些会		3	未采纳	7小时前
[心血管内科] 经常性睡觉前和起床后胸闷	3币	3	未采纳	7小时前
[冠心病] 冠心病，心悸气滞血瘀，可以吃新鲜虫草吗?	3币	3	未采纳	8小时前

Sample and Data Collection

We collected the patients' health-related questions on the crowdsourced health care information websites, and doctors provided their suggestions or advice in these websites. To examine the complete interaction process between the patient and the doctor, we chose questions that were already assigned as the best answer by the patients. We wrote a crawler in Python to download data in the crowdsourced health care information websites. For each user, we built a list of historical information, including the questions, answers, participating user identification numbers, and other features. The data were cleaned in advance by removing meaningless characters such as repeated characters and unanswered questions in text queries. Finally, 1524 complete interact process records were identified with 3245 answers in 2014-2015, and these were included in the empirical study.

Variables and Model Estimation

We tested our hypotheses by using the ordinary least squares (OLS) model. In our research, we chose the ability of the doctor who answered first as the cue of the central route and the patient's reward as the peripheral route information. We divided the abilities of the doctors into 2 categories: web-based ability (which refers to the average score given to the doctors by the patients based on their web-based medical service quality) and clinical ability (the professional title of the doctors in the hospital). Moreover, the question's complexity was included as the moderating variable.

Dependent Variables

1. Doctors' participation (D_Participation): For each question i , the number of doctors who answered was collected and its log value was used in the models. For each question, the doctors' answers were sorted by the answered date, and we captured all the service information about the doctor who answered first.

Independent Variables

Central Route Information

1. Title of the doctor who answered first (Dtitle_dummy): In China, doctors have titles that are evaluated by the government and the titles represents their clinical level of service in the hospitals; the different titles include the chief doctor, associate chief doctor, attending doctor, and others. We used 2 dummy variables to measure the doctors' titles: Dtitle_dummy1 and Dtitle_dummy2 (Figure 5).

2. Web-based ability of the doctor who answered first (D_Score): After receiving the doctors' web-based service, the patients can rate the doctors for the quality of their services, which ranges from 0 to 5 on the 120ask website.

Figure 5. Central route information.

$$Dtitle\_dummy1\begin{cases}=1, \text{ when doctor has a chief or associate chief title}\\ =0, \text{ others}\end{cases}$$

$$Dtitle\_dummy2\begin{cases}=1, \text{ when doctor has a attending title}\\ =0, \text{ other}\end{cases}$$

Peripheral Route Information

1. Reward assigned by the patient (P_Reward): Patients need to set a reward for their questions. According to the rule of the 120ask website, the maximum setting value of the reward is 100 CNY (1CNY= US $0.14).

Moderating Effects

1. Complexity of the patient's question (Q _Complexity): We used the length of the first doctor's reply to measure the complexity of the patient's question.

Control Variables

In our model, we also included other variables that could affect the doctors' behavior, namely, (1) patient's age (P_Age), (2) patient's gender (P_Gender), (3) time limit (P_ Deadline), (4) the response speed of the doctor who answered first (D_Reponse Speed), and (5) the total number of questions answered by the doctor who answered first (D_Assistance Number). We used these variables in our research model to control the effects of the 2 different routes on the doctors' participation. All variables and their descriptions are shown in Table 1.

Table 1. Description of the variables.

Variables	Variable symbol	Description
Dependent variable		
Doctors' participation	D_Participation	The number of doctors who answered the question *i*. Its log value is used in the models.
Independent variables		
Central route: the professional title of the doctor who answered first	Dtitle_dummy1 Dtitle_dummy2	The professional titles of the doctors represent their clinical abilities. Two dummy variables are used to measure doctor titles.
Central route: the web-based rating of the doctor who answered first	D_Score	The score that patients rate on the doctor's quality of medical services, which ranges from 0 to 5.
Peripheral route: reward	P_Reward	The reward that the patient assigns to the question *i*. The maximum value is 100 CNY[a].
Moderating effect		
Question's complexity	Q _Complexity	The number of characters in the first doctor's response is used to measure the complexity of the question that the patient posts. Its log value is used in the models.
Control variables		
Patient's age	P_Age	Its log value is used in the models.
Patient's gender	P_Gender	1 for male and 0 for female.
Time limit	P_ Deadline	The time limit that the patient sets to the questions. Its log value is used in the models.
Response speed of the doctor who answered first	D_ Response speed	The response speed of the doctor who answered first is included in the model.
Total number of questions by the doctor who answered first	D_ Assistance numbers	The total number of questions that the doctor has answered no matter whether he/she has received a reward.

[a] 1CNY= US $0.14

Model

The empirical model is as follows:

$$Ln(D\_Participants)=\beta_0 + \beta_1 P\_Gender + \beta_2 Ln(P\_Age) + \beta_3 Ln(P\_Deadline) + \beta_4 Ln(D\_Reponse\ speed) + \beta_5 D\_Assistance + \beta_{online} P\_Reward +$$

$$\beta_{offline1}Dtitle\_dummy1 + \beta_{offline2}Dtitle\_dummy2 + \beta_6 D\_Score + \beta_7 Ln(Q\_Complexity) + \beta_8 P\_Rewards \times Ln(Q\_Complexity) + \beta_9 Dtitle\_dummy1 \times Ln(Q\_Complexity) + \beta_{10} Dtitle\_dummy2 \times Ln(Q\_Complexity) + \beta_{11} D\_Score \times Ln(Q\_Complexity) + \varepsilon_0$$

Empirical Results

The summary of the statistics of our main variables and their correlations are presented in Table 2. All variables were correlated with the doctors' participation, except the patient's age and the total number of answers of the doctor who answered first. Meanwhile, the correlations between the independent variables and the control variables were low, which enabled us to obtain stable results.

The empirical results are shown in Table 3 hierarchically. The results for the model with the control variables are shown in Model 1, and then the independent variables, the moderating variable, and the interaction terms in Models 2-4 were added. The adjusted R^2 (>25%) and *F* values were reasonable and significant. All the variance inflation factor statistics for the variables were less than 2.0, which indicated the absence of multicollinearity.

We include control variables, that is, P_Gender, P_Age, P_Deadline, D_Response Speed, and D_Assistance Number, to address the potential endogenous issue. The results showed that all the control variables have correlations with doctor participation except P_Age and D_Assistance number. Our results revealed that when the poster (patient) is a male, the doctors' participation will increase (β=.114, *P*<.001). A longer deadline improves the doctors' participation significantly (β=.023, *P*=.006). Meanwhile, we also found that the response speed of the doctor who answered first positively influenced the following doctors' participation (β=.088, *P*<.001).

The central route hypothesis predicted that the ability of the doctor who answered first would have a positive effect on the following doctors' participation in the crowdsourced health care information websites. Based on the empirical results, we found that both web-based ability and the professional title of the doctor who answered first positively influenced the following doctors' participation, and the central route hypothesis was supported. With regard to the professional title of the doctor who answered first (in Model 2 of Table 3), the coefficients of Dtitle_dummy1 ($\beta_{offline1}$=.177, *P*<.001) and Dtitle_dummy2 ($\beta_{offline2}$=.063, *P*=.048) were positive and statistically significant. With regard to the professional title of the doctor who answered first (in Model 2 of Table 3), the coefficient of D_Score (β_{online}=.418, *P*<.001) was found to be positive and statistically significant. The peripheral cue hypothesis predicted that the reward has a positive effect on the doctors' participation in crowdsourced health care information websites. Based on the results in Table 3, we find that the peripheral cue hypothesis is supported based on the coefficient of the reward (β=.019, *P*<.001). Our model suggests that the question's complexity has a moderating effect on the relationships between the central route processing/peripheral route processing and doctors' participation. In Table 3, we find that the web-based ability of the doctor who answered first has a significant moderating effect (β=.186, *P*=.05), but the moderating effect of the professional titles of the doctors is not significant. We also obtained the opposite direction of the moderating effects of reward (β=–.0003, *P*=.10). Therefore, central route processing is partly supported and peripheral route processing is not supported.

Table 2. Descriptive statistics and correlations of the variables.

Variable, mean (SD)	D_Participation	P_Gender	P_Age	P_ Deadline	D_Reponse Speed	D_Assistance Number	P_Rewards	Dtitle_dummy1	Dtitle_dummy2	D_Score	Q_Complexity
D_Participation, 2.16 (1.228)											
r	1	0.114	–0.002	0.044	–0.037	0.294	0.341	0.183	–0.110	0.258	0.067
P value	—[a]	<.001	.94	.08	.15	<.001	<.001	<.001	<.001	<.001	.01
P_Gender, 0.46 (0.499)											
r	0.114	1	0.002	0.040	0.032	0.025	0.124	0.016	0	0.037	–0.021
P value	<.001	—	.95	.12	.22	.34	<.001	.54	.99	.15	.43
P_Age, 29.89 (16.62)											
r	–0.002	0.002	1	–0.008	–0.004	–0.031	0.040	–0.035	–0.008	0.011	–0.058
P value	.93	.95	—	.76	.89	.23	.13	.17	.75	.67	.02
P_ Deadline, 2.16 (1.228), 0.46 (0.499)											
r	0.044	0.040	–0.008	1	0.119	–0.039	–0.021	0.003	–0.151	0.120	0.144
P value	.08	.12	.76	—	<.001	.13	.41	.91	<.001	<.001	<.001
D_Reponse Speed, 173.50 (1457.6)											
r	–0.037	0.032	0	0.119	1	–0.052	–0.006	–0.014	0.009	–0.036	0.007
P value	.15	.22	.89	<.001	—	.04	.83	.59	.74	.16	.79
D_Assistance Number, 154.55 (122.6)											
r	0.294	0.025	–0.031	–0.039	–0.052	1	0.059	0.113	–0.129	0.295	0.008
P value	<.001	.34	.23	.13	.04	—	.02	<.001	<.001	<.001	.76
P_Rewards, 5.37 (8.76)											
r	0.341	0.124	0.040	–0.021	–0.006	0.059	1	0.020	0.032	0.025	0.168
P value	<.001	<.001	.13	.41	.83	.02	—	.44	.21	.34	<.001
Dtitle_dummy1, 0.26 (0.440)											
r	0.183	0.016	–0.035	0.003	–0.014	0.113	0.020	1	–0.544	0.278	0.056
P value	<.001	.54	.17	.91	.59	<.001	.44	—	<.001	<.001	.03
Dtitle_dummy2, 0.45 (0.498)											
r	–0.110	0	–0.008	–0.151	0.009	–0.129	0.032	–0.544	1	–0.370	–0.018
P value	<.001	.99	.75	<.001	.74	<.001	.21	<.001	—	<.001	.496
D_Score, 4.67 (0.201)											
r	0.258	0.037	0.011	0.120	–0.036	0.295	0.025	0.278	–0.370	1	0.017
P value	<.001	.15	.67	<.001	.16	<.001	.34	<.001	<.001	—	.52
Q_Complexity, 88.71 (64.48)											
r	0.067	–0.021	–0.058	0.144	0.007	0.008	0.168	0.056	–0.018	0.017	1
P value	.01	.43	.02	<.001	.79	.76	<.001	.03	.496	.52	—

[a] Not applicable.

Table 3. Empirical model results.

Variables	Model 1[a]		Model 2[b]		Model 3[c]		Model 4[d]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
P_Gender	.114 (.027)	<.001	.065 (.025)	.007	.065 (.025)	.008	.065 (.025)	.007
P_Age	.012 (.016)	.35	.015 (.015)	.32	.016 (.015)	.32	.015 (.015)	.27
P_ Deadline	.023 (.009)	<.001	.017 (.008)	.05	.021 (.008)	<.001	.023 (.008)	<.001
D_Response Speed	.088 (.007)	<.001	.067 (.007)	<.001	.064 (.007)	<.001	.065 (.007)	<.001
D_Assistance Number	–.035 (.024)	.22	–.025 (.022)	.32	–.022 (.022)	.32	–.19 (.022)	.28
P_Rewards	__[e]	—	.019 (.001)	<.001	.018 (.001)	<.001	.030 (.007)	<.001
Dtitle_dummy1	—	—	.177 (.034)	<.001	.186 (.034)	<.001	.307 (.173)	.06
Dtitle_dummy2	—	—	.063 (.032)	.05	.066 (.032)	.005	–.066 (.160)	.57
D_Score	—	—	.418 (.070)	<.001	.328 (.073)	<.001	–.386 (.307)	.11
Q_Complexity	—	—	—	—	.057 (.015)	<.001	–.807 (.368)	.049
P_Rewards×Q_Complexity	—	—	—	—	—	—	–.006 (.002)	<.001
Dtitle_dummy1×Q_Complexity	—	—	—	—	—	—	–.044 (.041)	.28
Dtitle_dummy2×Q_Complexity	—	—	—	—	—	—	.011 (.037)	.76
D_Score×Q_Complexity	—	—	—	—	—	—	.186 (.078)	.07

[a]Adjusted R^2: 0.104 ; *F* change: 34.083 (*P*<.001).

[b]Adjusted R^2: 0.244; *F* change: 66.852 (*P*<.001).

[c]Adjusted R^2: 0.251; *F* change: 14.030 (*P*<.001).

[d]Adjusted R^2: 0.253; *F* change: 2.076 (*P*=.08).

[e]Not available.

Robustness Check

To check the robustness of our results, we chose questions with a deadline of less than 41 days (the average value of P_ Deadline) as our sample. Finally, 1301 doctors were included in the model. A long deadline may reduce the doctors' enthusiasm to answer the questions as the payback time is unpredictable. In addition, the patients' sincerity may be questioned when they post questions with a long deadline. Table 4 presents the results of our model robustness, which was estimated using OLS. The results are consistent with our main findings, and our empirical results were found to be robust.

Table 4. Robustness check.

Variables	Model 1[a]		Model 2[b]		Model 3[c]		Model 4[d]	
	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value	β (SD)	*P* value
P_Gender	.117 (.028)	<.001	.072 (.026)	<.001	.073(.026)	<.001	.072 (.026)	.002
P_Age	.012 (.016)	.42	.015 (.015)	.39	.016(.015)	.40	.016 (.015)	.30
P_ Deadline	.022 (.011)	.02	.012 (.010)	.27	.021(.010)	.05	.023 (.010)	.003
D_Response Speed	.088 (.008)	<.001	.067 (.007)	<.001	.065(.007)	<.001	.066 (.007)	<.001
D_Assistance Number	–.038 (.024)	.17	–.026 (.022)	.36	–.024(.022)	.36	–.21(.022)	.35
P_Rewards	—[e]	—	.019 (.001)	<.001	.018(.001)	<.001	.031(.007)	<.001
Dtitle_dummy1	—	—	.180 (.035)	<.001	.192(.035)	<.001	.328 (.175)	.009
Dtitle_dummy2	—	—	.068 (.032)	.02	.072 (.032)	.02	–.070 (.162)	.66
D_Score	—	—	.414 (.070)	<.001	.321(.074)	<.001	–.393 (.310)	.13
Q_Complexity	—	—	—	—	.058(.015)	<.001	–.803 (.371)	<.001
P_Rewards×Q_Complexity	—	—	—	—	—	—	–.007 (.002)	<.001
Dtitle_dummy1×Q_Complexity	—	—	—	—	—	—	–.033 (.040)	.14
Dtitle_dummy2×Q_Complexity	—	—	—	—	—	—	.033 (.037)	.78
D_Score×Q_Complexity	—	—	—	—	—	—	.186 (.078)	.04

[a] Adjusted R^2: 0.105 ; *F* change: 33.492 (*P*<.001).

[b] Adjusted R^2: 0.244; *F* change: 64.256 (*P*<.001).

[c] Adjusted R^2: 0.251; *F* change: 13.692 (*P*<.001).

[d] Adjusted R^2: 0.253; *F* change: 2.179 (*P*=.09).

[e] Not available.

Discussion

Principal Findings

Overall, our results provide us with valuable insights into the role of the central and peripheral cues in crowdsourced health care information websites based on the framework of ELM. Our statistical evidence suggests that the following doctors' participation is related to the ability of the doctor who answered first. Based on the ELM, the central cues present the information needed for in-depth thinking. The ability of the doctor who answered first was used as the central cue in our study. In crowdsourced health care information websites, doctors could read the information in the prior answers before they answered the question. Based on the signal theory [61], we believe that the ability of the doctor who answered first can reflect the question's value and lead to a positive behavioral implication and increase the participation intention of the other doctors. Therefore, highly competent doctors should play a leading role in solving health problems.

The results relating to the rewards posted by the patients indicated that this variable is closely related to the intention of the doctors' participation. Our results suggest that similar to other crowdsourcing fields [25,51], doctors are very concerned about the reward. The reward drives doctors to participate in services that help patients solve their health problems. We also believe that more doctors will participate in providing medical services through crowdsourced health care information websites when the reward is higher than they expected. Therefore, setting a high reward can increase the participation of a large number of doctors to answer the question.

Our empirical results show that there is a significant moderating effect between the question complexity and the dual route. For the central route, our results show that the question's complexity can enhance the effect of the ability of the doctor who answered first on the following doctors' participation. Questions with high complexity are often more worthy for doctors with high competencies to answer, and our results suggest that highly competent doctors should take the responsibility to solve questions with high complexity. However, we found that question complexity does not have a significant moderating effect. A possible reason is that the doctor's web-based ability (the average score rated by the patients) represents the doctor's comprehensive web-based ability, which is more effective than the professional title in crowdsourced health care information websites because the entire interaction process is performed on the internet. Another reason is that the active group of doctors in crowdsourced health care information websites is mostly middle-level doctors who are younger and have more time to help patients in web-based medical services. For the peripheral route, we found that the reward was not very important when the question complexity was high because of the doctor's altruism, which has also been verified in other online health communities [29]. Therefore, the question complexity positively moderates the influence of the ability of the first responding doctor on the following doctors' participation and negatively moderates the influence of the reward factor on the following doctors' participation.

Contributions of This Study

This study has made the following contributions. First, to our knowledge, we are among the first to extend the ELM model to crowdsourced health care information websites. Previous studies have often used ELM in consumer adoption or satisfaction, information technology adoption, information adoption, and other areas [62-64]. We extended these previous studies by using ELM to investigate the doctors' participation in providing medical services through crowdsourced health care information websites and explored the different routes of different cues for understanding the behaviors of doctors' participation in crowdsourced health care information websites. Moreover, we used the question's complexity to investigate the moderating effects on the roles of the 2 routes. Second, our study has added valuable information to the existing studies on online health communities. The previous studies mainly focused on one-to-one consultant service to study the satisfaction of the patient or to study the relationship between the signals of other doctors (eg, price and reputation) and patient's choice [38,65]. Our study investigated the doctors' participation in one-to-crowd crowdsourced health care information websites. Third, we focused on expert-based question-and-answer websites, whereas the existing studies are based on Baidu Zhidao and Wiki Answers, which belongs to the ordinary community-based question-and-answer websites [66,67]. Our study broadens the research on expert-based question-and-answer websites, especially in the medical domain.

Our research has three major implications for practice. First, we found that the behaviors of the doctors involved in answering the patient's questions are influenced by the behavior of the ability of the doctor who first responded to the patient's question. Therefore, competent doctors should be encouraged to take up the leadership positions in online health communities and be actively involved in crowdsourced health care information websites. Second, according to our results, if patients want to receive more answers, they should increase the rewards for the question or invite highly competent doctors to answer the questions. Third, if the managers of the online health communities want to operate the platform successfully and make a profit, they should encourage doctors by providing an incentive mechanism to answer the question quickly and thoughtfully, as shown in our results.

Limitations of This Study

This study had the following limitations. First, this study selected only 1 online health community to investigate the participation behaviors of the doctors. Future studies should select different online health communities to compare the differences. Second, future research should consider other types of questions of the patients, such as questions related to emotional support needs or professional health care needs. Third, future research should adopt a longitudinal perspective to overcome the disadvantages of the cross-sectional data and explore the dynamics in the relationships as well.

Conclusion

This research explored the effects of the ability of the doctor who answered patients' questions first as well as the effects of rewards on the following doctors' participation in crowdsourced health care information websites. We also investigated the moderating effects of the question's complexity on these relationships. We developed a mathematical model to test our hypotheses. The empirical results supported most of our hypotheses. This study can help academicians to better understand the evaluation and the decision processes used by doctors when considering the web-based health-related crowdsourcing services. Moreover, this study has provided several implications for the practice of online health community managers and users.

Conflicts of Interest

None declared.

References

1. Scheffler R. Forecasting the global shortage of physicians: an economic- and needs-based approach. Bull World Health Organ 2008 Jul 01;86(7):516-523. [doi: 10.2471/blt.07.046474]
2. Zhang X, Wen D, Liang J, Lei J. How the public uses social media wechat to obtain health information in china: a survey study. BMC Med Inform Decis Mak 2017 Jul 5;17(S2). [doi: 10.1186/s12911-017-0470-0]
3. Cao W, Zhang X, Xu K, Wang Y. Modeling Online Health Information-Seeking Behavior in China: The Roles of Source Characteristics, Reward Assessment, and Internet Self-Efficacy. Health Commun 2016 Sep 09;31(9):1105-1114. [doi: 10.1080/10410236.2015.1045236] [Medline: 26861963]
4. Wu H, Lu N. Online written consultation, telephone consultation and offline appointment: An examination of the channel effect in online health communities. Int J Med Inform 2017 Nov;107:107-119. [doi: 10.1016/j.ijmedinf.2017.08.009] [Medline: 29029686]

5. iResearch. URL: http://www.iresearchchina.com/ [accessed 2020-06-09]
6. Matsumoto M, Inoue K, Kashima S, Takeuchi K. Does the insufficient supply of physicians worsen their urban-rural distribution? A Hiroshima-Nagasaki comparison. Rural Remote Health 2012;12:2085 [FREE Full text] [Medline: 22533349]
7. Sen K, Ghosh K. Designing Effective Crowdsourcing Systems for the Healthcare Industry. Crowdsourcing: Concepts, Methodologies, Tools, and Applications. PA: IGI Global; 2019. URL: https://scholarspace.manoa.hawaii.edu/bitstream/10125/41556/paper0407.pdf [accessed 2020-05-29]
8. 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]
9. Ghezzi A, Gabelloni D, Martini A, Natalicchio A. Crowdsourcing: A Review and Suggestions for Future Research. International Journal of Management Reviews 2017 Jan 19;20(2):343-363. [doi: 10.1111/ijmr.12135]
10. Lee H, Seo S. What Determines an Agreeable and Adoptable Idea? A Study of User Ideas on MyStarbucksIdea.com. 2013 Presented at: 46th Hawaii International Conference on System Sciences; 2013; Wailea, Maui, HI, USA p. 3207-3217. [doi: 10.1109/HICSS.2013.604]
11. Lossio-Ventura JA, Hogan W, Modave F, Guo Y, He Z, Yang X, et al. OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):55 [FREE Full text] [doi: 10.1186/s12911-018-0635-5] [Medline: 30066655]
12. Leal-Neto OB, Dimech GS, Libel M, Oliveira W, Ferreira JP. Digital disease detection and participatory surveillance: overview and perspectives for Brazil. Rev. Saúde Pública 2016;50. [doi: 10.1590/s1518-8787.2016050006201]
13. Doan A, Ramakrishnan R, Halevy AY. Crowdsourcing systems on the World-Wide Web. Commun. ACM 2011 Apr;54(4):86-96. [doi: 10.1145/1924421.1924442]
14. Brabham DC. Crowdsourcing as a Model for Problem Solving. Convergence 2008 Feb;14(1):75-90. [doi: 10.1177/1354856507084420]
15. Brabham DC. Crowdsourcing as a Model for Problem Solving. Convergence 2008 Feb;14(1):75-90. [doi: 10.1177/1354856507084420]
16. Zhao Y, Zhu Q. Evaluation on crowdsourcing research: Current status and future direction. Inf Syst Front 2012 Apr 11;16(3):417-434. [doi: 10.1007/s10796-012-9350-4]
17. Wójcik OP, Brownstein JS, Chunara R, Johansson MA. Public health for the people: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol 2014;11:7 [FREE Full text] [doi: 10.1186/1742-7622-11-7] [Medline: 24991229]
18. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res 2009 Mar 27;11(1):e11 [FREE Full text] [doi: 10.2196/jmir.1157] [Medline: 19329408]
19. Li Y, Du N, Liu C, Xie Y, Fan W, Li Q, et al. Reliable Medical Diagnosis from Crowdsourcing: Discover Trustworthy Answers from Non-Experts. 2017 Presented at: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining; 2017; New York, United States p. 253-261. [doi: 10.1145/3018661.3018688]
20. Li J, Liu H, Zhang Y, Xing C. A Health QA with Enhanced User Interfaces. In: 2016 13th Web Information Systems and Applications Conference (WISA). 2016 Presented at: 13th Web Information Systems and Applications Conference (WISA); 23-25 Sept. 2016; Wuhan, China p. 173. [doi: 10.1109/wisa.2016.43]
21. Yin Y, Zhang Y, Liu X, Zhang Y, Xing C, Chen H. HealthQA: A Chinese QA Summary System for Smart Health. In: International Conference on Smart Health. 2014 Presented at: International Conference on Smart Health; July 10-11, 2014; Beijing, China p. 51-62. [doi: 10.1007/978-3-319-08416-9_6]
22. Jeff H. The rise of crowdsourcing. Wired magazine. 2011. URL: https://www.wired.com/2006/06/crowds/ [accessed 2020-05-05]
23. Saxton GD, Oh O, Kishore R. Rules of Crowdsourcing: Models, Issues, and Systems of Control. Information Systems Management 2013 Jan;30(1):2-20. [doi: 10.1080/10580530.2013.739883]
24. Loukis E, Charalabidis Y. Active and Passive Crowdsourcing in Government. Policy Practice and Digital Science 2015:261-289. [doi: 10.1007/978-3-319-12784-2_12]
25. Harris CG. Youre Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. 2011 Jan. URL: https://www.researchgate.net/publication/228566973_Youre_Hired_An_Examination_of_Crowdsourcing_Incentive_Models_in_Human_Resource_Tasks [accessed 2020-05-05]
26. Lease M, Carvalho VR, Yilmaz E. Crowdsourcing for search and data mining. SIGIR Forum 2011 May 24;45(1):18-24. [doi: 10.1145/1988852.1988856]
27. Kittur A, Chi EH, Suh B. Crowdsourcing user studies with Mechanical Turk. 2008 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2008; Florence Italy p. 453-456. [doi: 10.1145/1357054.1357127]
28. Guan T, Wang L, Jin J, Song X. Knowledge contribution behavior in online Q&A communities: An empirical investigation. Computers in Human Behavior 2018 Apr;81:137-147. [doi: 10.1016/j.chb.2017.12.023]

29. Nam KK, Ackerman MS, Adamic LA. Questions in, knowledge in? a study of naver's question answering community. 2009 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2009; New York, United States p. 779-788. [doi: 10.1145/1518701.1518821]
30. Hsu M, Ju TL, Yen C, Chang C. Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. International Journal of Human-Computer Studies 2007 Feb;65(2):153-169. [doi: 10.1016/j.ijhcs.2006.09.003]
31. Silberman MS, Irani L, Ross J. Ethics and tactics of professional crowdwork. XRDS:Crossroads, The ACM Magazine for Students 2010 Dec;17(2):39-43. [doi: 10.1145/1869086.1869100]
32. Huberman BA, Romero DM, Wu F. Crowdsourcing, attention and productivity. Journal of Information Science 2009 Oct 09;35(6):758-765. [doi: 10.1177/0165551509346786]
33. Petty R, Cacioppo J. The Elaboration Likelihood Model of Persuasion. Communication and Persuasion 1986:123. [doi: 10.1016/s0065-2601(08)60214-2]
34. Tam KY, Ho SY. Web Personalization as a Persuasion Strategy: An Elaboration Likelihood Model Perspective. Information Systems Research 2005 Sep;16(3):271-291. [doi: 10.1287/isre.1050.0058]
35. Zhou T. Understanding users’ initial trust in mobile banking: An elaboration likelihood perspective. Computers in Human Behavior 2012 Jul;28(4):1518-1525. [doi: 10.1016/j.chb.2012.03.021]
36. Petty RE, Cacioppo JT, Goldman R. Personal involvement as a determinant of argument-based persuasion. Journal of Personality and Social Psychology 1981;41(5):847-855. [doi: 10.1037/0022-3514.41.5.847]
37. Bi S, Liu Z, Usman K. The influence of online information on investing decisions of reward-based crowdfunding. Journal of Business Research 2017 Feb;71:10-18. [doi: 10.1016/j.jbusres.2016.10.001]
38. Cao X, Liu Y, Zhu Z, Hu J, Chen X. Online selection of a physician by patients: Empirical study from elaboration likelihood perspective. Computers in Human Behavior 2017 Aug;73:403-412. [doi: 10.1016/j.chb.2017.03.060]
39. Park D, Lee J, Han I. The Effect of On-Line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement. International Journal of Electronic Commerce 2014 Dec 08;11(4):125-148. [doi: 10.2753/jec1086-4415110405]
40. Cheung CMK, Lee MKO, Rabjohn N. The impact of electronic word-of-mouth. Internet Research 2008 Jun 06;18(3):229-247. [doi: 10.1108/10662240810883290]
41. Cheung C, Sia C, Kuan K. Is This Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an ELM Perspective. JAIS 2012 Aug;13(8):618-635. [doi: 10.17705/1jais.00305]
42. Duan, Gu, Whinston. Informational Cascades and Software Adoption on the Internet: An Empirical Investigation. MIS Quarterly 2009;33(1):23. [doi: 10.2307/20650277]
43. Simpson PM, Siguaw JA, Cadogan JW. Understanding the consumer propensity to observe. European Journal of Marketing 2008 Feb 15;42(1/2):196-221. [doi: 10.1108/03090560810840970]
44. Cheung CM, Xiao BS, Liu IL. Do actions speak louder than voices? The signaling role of social information cues in influencing consumer purchase decisions. Decision Support Systems 2014 Sep;65:50-58. [doi: 10.1016/j.dss.2014.05.002]
45. Lee J, Park D, Han I. The effect of negative online consumer reviews on product attitude: An information processing view. Electronic Commerce Research and Applications 2008 Sep;7(3):341-352. [doi: 10.1016/j.elerap.2007.05.004]
46. Borst I. Understanding Crowdsourcing: Effects of motivation and rewards on participation and performance in voluntary online activities. 2010. URL: https://www.erim.eur.nl/research/news/detail/1703-understanding-crowdsourcing-effects-of-motivation-and-rewards-on-participation-and-performance-in-voluntary-online-activities/ [accessed 2020-05-01]
47. Sun Y, Wang N, Yin C, Zhang JX. Understanding the relationships between motivators and effort in crowdsourcing marketplaces: A nonlinear analysis. International Journal of Information Management 2015 Jun;35(3):267-276. [doi: 10.1016/j.ijinfomgt.2015.01.009]
48. Ye H, Kankanhalli A. Solvers’ participation in crowdsourcing platforms: Examining the impacts of trust, and benefit and cost factors. The Journal of Strategic Information Systems 2017 Jun;26(2):101-117. [doi: 10.1016/j.jsis.2017.02.001]
49. Lai H, Chen TT. Knowledge sharing in interest online communities: A comparison of posters and lurkers. Computers in Human Behavior 2014 Jun;35:295-306. [doi: 10.1016/j.chb.2014.02.004]
50. Yang D, Xue G, Fang X, Tang J. Incentive Mechanisms for Crowdsensing: Crowdsourcing With Smartphones. IEEE/ACM Trans. Networking 2016 Jun;24(3):1732-1744. [doi: 10.1109/tnet.2015.2421897]
51. Moreno A, Rosa J, Szymanski B. Reward System for Completing FAQs. 2009 Presented at: Conference on Artificial Intelligence Research and Development: Proceedings of the International Conference of the Catalan Association for Artificial Intelligence; 2009; Cardona, Spain p. 361-370 URL: https://www.researchgate.net/publication/221045373_Reward_System_for_Completing_FAQs
52. Horton J, Chilton L. The labor economics of paid crowdsourcing. In: Proceedings of the 11th ACM conference on Electronic commerce. 2010 Presented at: Proceedings of the 11th ACM conference on Electronic commerce; June 2010; Cambridge Massachusetts USA p. 209. [doi: 10.1145/1807342.1807376]
53. Kazai G. An Exploration of the Influence that Task Parameters have on the Performance of Crowds. 2010 Presented at: Proceedings of the Crowdconf; 2010; Cork, Ireland.

54. Meservy TO, Jensen ML, Fadel KJ. Evaluation of Competing Candidate Solutions in Electronic Networks of Practice. Information Systems Research 2014 Mar;25(1):15-34. [doi: 10.1287/isre.2013.0502]
55. Bhattacherjee A, Sanford C. Influence Processes for Information Technology Acceptance: An Elaboration Likelihood Model. MIS Quarterly 2006;30(4):805. [doi: 10.2307/25148755]
56. Petty R, Cacioppo J. The Ability to Elaborate in a Relatively Objective Manner. In: Communication and Persuasion. New York: Springer; 1986.
57. Petty R, Cacioppo J. Central and Peripheral Routes to Attitude Change. In: Communication and Persuasion. New York: Springer; 2016.
58. Wu H, Lu N. Service provision, pricing, and patient satisfaction in online health communities. Int J Med Inform 2018 Feb;110:77-89. [doi: 10.1016/j.ijmedinf.2017.11.009] [Medline: 29331257]
59. Zhang M, Wu T, Guo X, Liu X, Sun W. The Effects of the Externality of Public Goods on Doctor's Private Benefit: Evidence from Online Health Community. 2017 Presented at: International Conference on Smart Health; 2017; Hong Kong, China p. 149-160. [doi: 10.1007/978-3-319-67964-8_14]
60. China Online Medical Industry Data Monitoring Report. 2016. URL: https://www.iresearch.com.cn/Detail/report?id=2551&isfree=0 [accessed 2020-06-09]
61. Mavlanova T, Benbunan-Fich R, Koufaris M. Signaling theory and information asymmetry in online commerce. Information & Management 2012 Jul;49(5):240-247. [doi: 10.1016/j.im.2012.05.004]
62. Ghose A, Ipeirotis PG, Li B. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. Marketing Science 2012 May;31(3):493-520. [doi: 10.1287/mksc.1110.0700]
63. Archak N, Ghose A, Ipeirotis PG. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. Management Science. 2011. URL: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1150&context=oid_papers [accessed 2020-05-03]
64. Yin D, Mitra S, Zhang H. Research Note—When Do Consumers Value Positive vs. Negative Reviews? An Empirical Investigation of Confirmation Bias in Online Word of Mouth. Information Systems Research 2016 Mar;27(1):131-144. [doi: 10.1287/isre.2015.0617]
65. Lu N, Wu H. Exploring the impact of word-of-mouth about Physicians' service quality on patient choice based on online health communities. BMC Med Inform Decis Mak 2016 Nov 26;16(1):151 [FREE Full text] [doi: 10.1186/s12911-016-0386-0] [Medline: 27888834]
66. Chen C, Wu K, Srinivasan V, Bharadwaj R. The best answers? Think twice: Online detection of commercial campaigns in the CQA forums. 2013 Presented at: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013); 2013; Niagara Falls, ON p. 458-465. [doi: 10.1145/2492517.2492553]
67. Liu Z, Jansen BJ. Factors influencing the response rate in social question and answering behavior. 2013 Presented at: Proceedings of the 2013 conference on Computer supported cooperative work; 2013; San Antonio Texas USA p. 1263-1274. [doi: 10.1145/2441776.2441918]

Abbreviations

ELM: elaboration-likelihood model
OLS: ordinary least squares

Edited by C Lovis; submitted 15.10.19; peer-reviewed by W Song, C Zhang, O Leal Neto; comments to author 26.10.19; revised version received 16.11.19; accepted 01.02.20; published 29.06.20.

Please cite as:
Si Y, Wu H, Liu Q
Factors Influencing Doctors' Participation in the Provision of Medical Services Through Crowdsourced Health Care Information Websites: Elaboration-Likelihood Perspective Study
JMIR Med Inform 2020;8(6):e16704
URL: http://medinform.jmir.org/2020/6/e16704/
doi:10.2196/16704
PMID:32597787

Original Paper

Detecting and Filtering Immune-Related Adverse Events Signal Based on Text Mining and Observational Health Data Sciences and Informatics Common Data Model: Framework Development Study

Yue Yu[1], PhD; Kathryn Ruddy[2], MD; Aaron Mansfield[2], MD; Nansu Zong[1], PhD; Andrew Wen[1], MSc; Shintaro Tsuji[1], PhD; Ming Huang[1], PhD; Hongfang Liu[1], PhD; Nilay Shah[1], PhD; Guoqian Jiang[1], MD, PhD

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

[2]Division of Medical Oncology, Department of Oncology, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Guoqian Jiang, MD, PhD
Department of Health Sciences Research
Mayo Clinic
200 First St SW
Rochester, MN, 55905
United States
Phone: 1 507 284 2511
Email: jiang.guoqian@mayo.edu

Abstract

Background: Immune checkpoint inhibitors are associated with unique immune-related adverse events (irAEs). As most of the immune checkpoint inhibitors are new to the market, it is important to conduct studies using real-world data sources to investigate their safety profiles.

Objective: The aim of the study was to develop a framework for signal detection and filtration of novel irAEs for 6 Food and Drug Administration–approved immune checkpoint inhibitors.

Methods: In our framework, we first used the Food and Drug Administration’s Adverse Event Reporting System (FAERS) standardized in an Observational Health Data Sciences and Informatics (OHDSI) common data model (CDM) to collect immune checkpoint inhibitor-related event data and conducted irAE signal detection. OHDSI CDM is a standard-driven data model that focuses on transforming different databases into a common format and standardizing medical terms to a common representation. We then filtered those already known irAEs from drug labels and literature by using a customized text-mining pipeline based on clinical text analysis and knowledge extraction system with Medical Dictionary for Regulatory Activities (MedDRA) as a dictionary. Finally, we classified the irAE detection results into three different categories to discover potentially new irAE signals.

Results: By our text-mining pipeline, 490 irAE terms were identified from drug labels, and 918 terms were identified from the literature. In addition, of the 94 positive signals detected using CDM-based FAERS, 53 signals (56%) were labeled signals, 10 (11%) were unlabeled published signals, and 31 (33%) were potentially new signals.

Conclusions: We demonstrated that our approach is effective for irAE signal detection and filtration. Moreover, our CDM-based framework could facilitate adverse drug events detection and filtration toward the goal of next-generation pharmacovigilance that seamlessly integrates electronic health record data for improved signal detection.

(JMIR Med Inform 2020;8(6):e17353) doi:10.2196/17353

KEYWORDS

immunotherapy/adverse effects; drug-related side effects and adverse reactions; pharmacovigilance; adverse drug reaction reporting systems/standards; text mining

Introduction

Background

Immunotherapy activates a patient's immune system for therapeutic benefit against cancer [1]. One type of immunotherapy, immune checkpoint inhibition, has recently been found to be promising for the treatment of certain types of cancer. Immune checkpoint inhibitors can block negative regulators (checkpoints) of T-cell function that exist on both immune and tumor cells. This blockage could enhance antitumor immunity by allowing T cells to kill cancer cells [2]. Notably, the Nobel Prize in Physiology or Medicine in 2018 was awarded to James Allison and Tasuku Honjo for their work on immune checkpoint inhibitors [3]. From 2011 to 2017, the US Food and Drug Administration (FDA) has approved a total of 6 immune checkpoint–directed antibodies for the treatment of specific tumors. By increasing the activity of the immune system, immune checkpoint inhibitors can have inflammatory side effects, which are often termed as immune-related adverse events (irAEs) [4]. The most recognized irAEs include dermatitis, colitis, hepatitis, pancreatitis, pneumonitis, and hypophysitis [5]. IrAEs are mostly of mild to moderate severity, but at times, these can be serious, irreversible, or even fatal. Nevertheless, several studies have indicated that immune checkpoint inhibitors have a better safety profile than many traditional chemotherapies [6-8]. As these immune checkpoint inhibitor agents are new to the market, investigation of their safety profiles in real-world practice is critical [4]. Traditionally, one of the most important ways to detect postmarketing safety profiles of drugs is to conduct pharmacovigilance studies using a spontaneous reporting system (SRS) database [9]. SRS is a system whereby case reports of adverse drug events (ADEs) are voluntarily submitted by health professionals and pharmaceutical companies to the national pharmacovigilance center [10]. Several studies have focused on the irAEs post marketing pharmacovigilance through the analysis of an SRS database such as the US Food and Drug Administration's Adverse Event Reporting System (FAERS) or World Health Organization (WHO)'s VigeBase [11-15].

Although there have been some previous studies that utilized SRS to detect irAEs, it is still essential to investigate new irAE signals to help the research community recognize a comprehensive drug safety profile for these immune checkpoint antibodies. However, it is now also recognized that traditional SRS-based ADE detection methods only focus on detecting statistically significant drug-event pairs from the SRS database, and these methods often face challenges in identifying those *new* pharmacovigilance signals automatically. Hauben and Aronson [16] proposed a widely used definition of a drug safety surveillance signal. This definition is also issued by the Council for International Organizations of Medical Sciences [17], an international organization established jointly by WHO and the United Nations Educational, Scientific and Cultural Organization, which is famous for establishing guidelines for international pharmacovigilance. According to this definition, a pharmacovigilance signal "represents an association that is new and important, or a new aspect of a known association, and has not been previously investigated and refuted." We can note that a detected drug-event association that is not fully recognized by the previous investigation could be seen as a *signal* in this definition. These new signals can be considered to be valuable starting points for further investigation and validation. However, for a current large-scale FAERS-based pharmacovigilance study, most of the detected drug-event associations are already recognized by the existing knowledge. Some of these signals have been discovered by clinical trials before approval or by the postmarketing pharmacovigilance study [18]. To filter the known drug-event associations, health care professionals often have to manually review a substantial number of drug safety-related texts, such as drug labels or biomedical literature, to determine whether these novel ADE signals are worthy of further validation [19-21]. It is typically laborious and imprecise to manually review these ADE signals, despite the promising results achieved by existing studies, such as those by Xu et al [22,23] and Yeleswarapu et al [24]. However, these studies have focused on ranking and finding the most significant detection results and did not consider identification of novel ADE signals, which are worth further investigation. Text-mining methods allow for a more efficient way to filter the drug-event associations that are already known by existing knowledge, by extracting known ADEs from drug labels and literature. By automatically filtering out existing signals, this effort can not only discover novel irAE signals detected by the FAERS but may also reduce the labor involved in human intervention.

Objectives

The objective of this study was to develop a framework for novel signal detection and filtration of irAEs. First, we normalized the FAERS using the Observational Health Data Sciences and Informatics (OHDSI) CDM to improve data standardization and quality to facilitate data collection and analysis. To detect irAEs, we selected all the 6 immune checkpoint inhibitors approved by the FDA before 2018 as our research object. We collected all standardized adverse event data regarding the 6 immune checkpoint inhibitors. Then, the reporting odds ratio (ROR) is utilized to detect the irAEs signal. To filter out those already known irAEs, a customized text-mining pipeline is implemented using clinical text analysis and knowledge extraction system (cTAKES) with MedDRA as a dictionary. Finally, we classified the irAE detection results into three different categories, including potentially new irAE signals.

Methods

Materials

Food and Drug Administration's Adverse Event Reporting System

The FAERS [25] is a database that contains information on adverse event and medication error reports submitted to FDA. FAERS is designed to support postmarketing safety surveillance. All voluntary adverse event reports in FAERS could be submitted by health care professionals (such as physicians, pharmacists, nurses, and others), consumers (such as patients, family members, lawyers, and others), and manufacturers. There are 7 tables in the FAERS database, which includes patient demographic table, drug table, adverse reaction table, patient outcome table, report source table, drug therapy table, and indication table. FAERS will update quarterly and can be downloaded from the FDA website. The adverse event name in FAERS is standardized by MedDRA, a rich and highly specific standardized medical terminology. However, the drug name in FAERS is not standard, which may be a drug ingredient name, a brand name, a clinical drug component, or even a spelling error. Some other information such as drug unit and drug dosage are also nonstandard. Therefore, it is important to conduct the data preprocessing to normalize the data in FAERS before the implementation of adverse event signal detection. In this study, we used the FAERS data covering the period from September 2012 to March 2017.

Observational Health Data Sciences and Informatics Common Data Model

The OHDSI common data model (CDM) [26], also known as Observational Medical Outcomes Partnership CDM, is a data model designed for the systematic analysis of disparate observational databases. OHDSI CDM focuses on transforming different observational databases into a common format (data model) and a common representation (terminologies, vocabularies, coding schemes). As of February 23, 2108, version V5.3 of the CDM was released, containing 37 tables in 6 categories: standardized clinical data, standardized health system data, standardized health economics, standardized metadata, standardized vocabularies, and standardized derived elements. In fact, terminology normalization enabled by standard vocabularies with a focus on systematized nomenclature of medicine-clinical terms (SNOMED CT), logical observation identifiers names and codes, and RxNorm is a strong characteristic of the OHDSI CDM. One of the advantages of using a CDM-based database to conduct a pharmacovigilance study is that we could build a standard query as the same standard terminologies are utilized to represent the medical concepts across the different observational databases. This allows for collaborative pharmacovigilance research across different institutions.

Food and Drug Administration Drug Label

We searched the DailyMed website to collect the drug labels of 6 FDA-approved immune checkpoint inhibitors [27]. DailyMed, developed and maintained by the National Library of Medicine, is the official provider of FDA drug label information. We downloaded the drug labels of the 6 immune checkpoint inhibitors in January 2018. These drug labels were downloaded in the structured product labeling (SPL) format, which is a document markup standard approved by Health Level Seven (HL7) and adopted by the FDA as a mechanism for exchanging product and facility information. We extracted the text under the section WARNINGS AND PRECAUTIONS and the section ADVERSE REACTIONS from the SPL files of 6 labels as the dataset of drug label text mining.

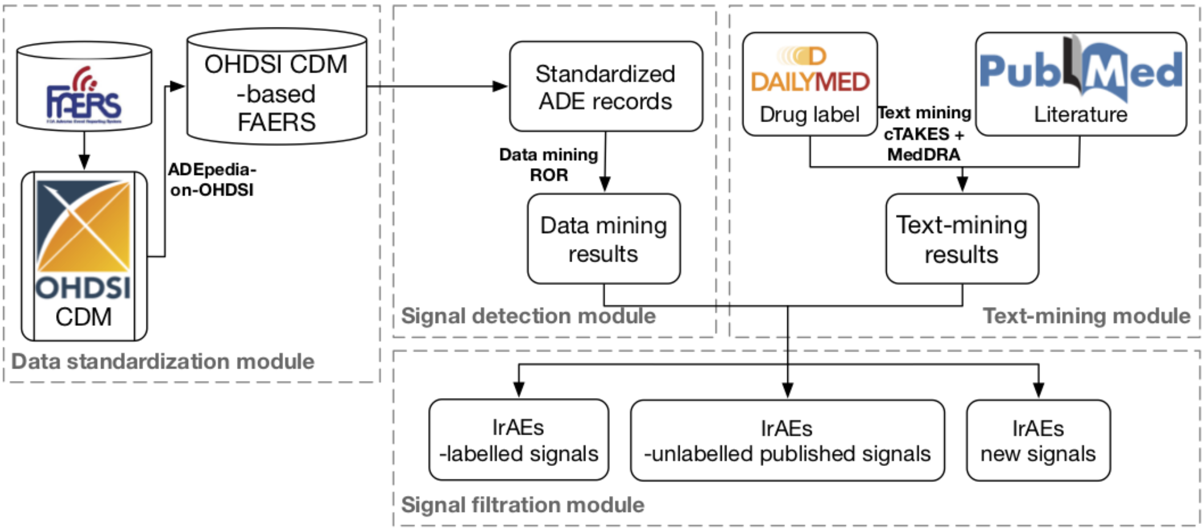
Immune-Related Adverse Events–Related PubMed Literature

We retrieved literature from PubMed [28] and built an irAE-related literature text-mining dataset. The query “immune-related [All Fields] AND adverse [All Fields] AND events [All Fields]” (retrieve date: January 2018) was used to retrieve literature from PubMed. A total of 679 irAEs-related literature was obtained. Then, we downloaded the abstract of 679 papers and the full text of 20 review articles as the irAE-related literature text-mining dataset.

Methods

Using FAERS standardized in the OHDSI CDM, we developed a framework for signal detection and filtration of irAEs, as shown in Figure 1. The framework mainly contains 4 modules as follows (data standardization module, signal detection module, text-mining module, and signal filtration module).

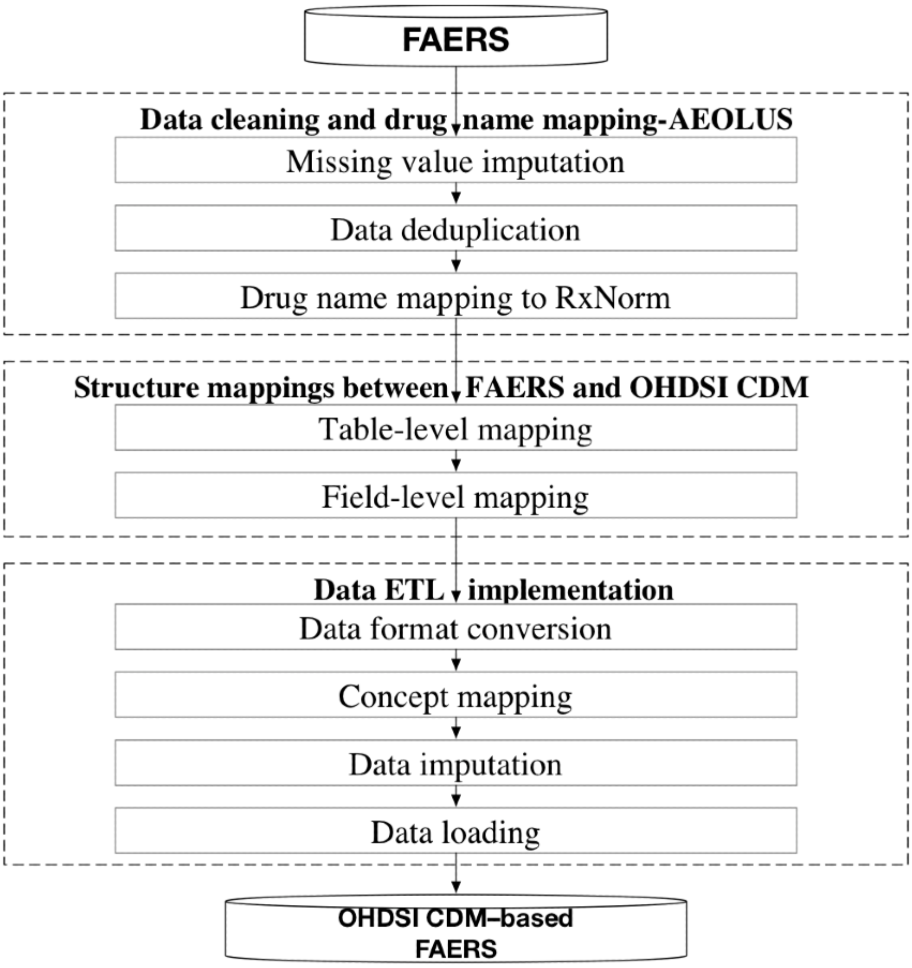
Figure 1. System architecture of our standards-driven framework. ADE: adverse drug events; CDM: common data model; cTAKES: clinical text analysis and knowledge extraction system; FAERS, Food and Drug Administration’s adverse event reporting system; IrAEs: immune-related adverse events; MedDRA: medical dictionary for regulatory activities; OHDSI: observational health data sciences and informatics; ROR: reporting odds ratio.



Data Standardization Module

In FAERS, some data are nonstandard. For example, a drug name in FAERS might be a drug ingredient name, a brand name, a clinical drug component, or even a spelling error. This data standardization problem would cause inconvenience in data collection and integration and introduce bias in data analysis. In this study, we developed a next-generation pharmacovigilance signal detection platform, ADEpedia-on-OHDSI [29], to standardize FAERS and integrate it with electronic health record (EHR) data by OHDSI CDM. Specifically, we used and extended adverse event open learning through universal standardization (AEOLUS)—an integration process developed by Banda et al [30] to develop an extract, transform, and load (ETL) process to transform FAERS data into OHDSI CDM. AEOLUS focuses on building a standard process for FAERS data deduplication and tooling for mapping drug names to RxNorm concepts and outcomes to SNOMED CT concepts. We further developed an ETL tool to convert FAERS into OHDSI CDM by data structure mapping, medical concept mapping, and data imputation. The 3-step ETL process of the ADEpedia-on-OHDSI platform is shown in Figure 2. (1) Data cleaning and drug name mapping: we used AEOLUS to conduct data deduplication and drug name mapping. (2) Structure mappings between FAERS schema and OHDSI CDM schema: we created structure mappings by choosing appropriate tables or fields between the OHDSI CDM and FAERS. (3) Data ETL implementation: we designed different ETL strategies and then loaded the raw FAERS data into the OHDSI CDM. More details about the ETL process of the ADEpedia-on-OHDSI platform can be found in our published paper [29]. After the ETL process, all the standardized FAERS data were stored in the relational database in the OHDSI CDM format. In this study, we utilized pgAdmin 4 (The pgAdmin Development Team) to operate and maintain our ADEpedia-on-OHDSI platform.

Figure 2. Extract, transform, and load process of converting Food and Drug Administration's Adverse Events Reporting System into Observational Health Data Sciences and Informatics common data model. AEOLUS, adverse event open learning through universal standardization; CDM, common data model; ETL: extract, transform, and load; FAERS, Food and Drug Administration's adverse event reporting system; OHDSI, observational health data sciences and informatics.



Signal Detection Module

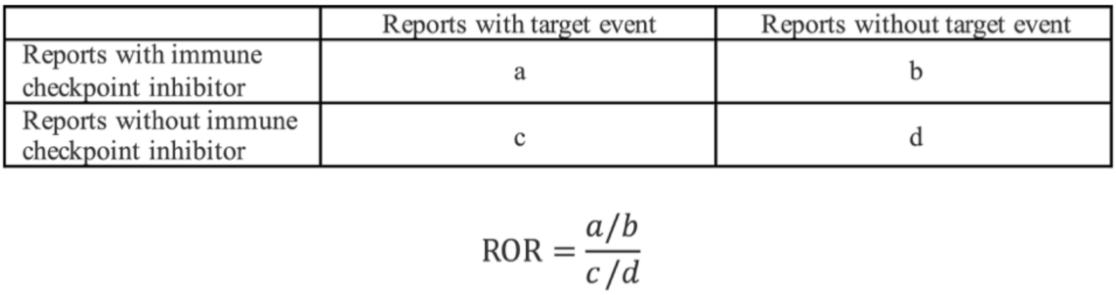
In this research, we define the adverse drug event (ADE) *signal* as the significant drug-adverse event associations detected by the detection algorithm using FAERS data. We implemented signal detection algorithms to detect potential irAE signals related to the 6 immune checkpoint inhibitor drugs approved by the FDA (ie, ipilimumab, pembrolizumab, nivolumab, atezolizumab, durvalumab, and avelumab). The active ingredient drug name, brand name, and the corresponding standard concept of those drugs are shown in Table 1. In addition, to facilitate the collection of irAE reports from our CDM-based FAERS, we built a standard query by checking all the synonyms of the 6 standard drug concepts in OHDSI ATHENA standardized vocabularies [31]. OHDSI concept_id related to all the ingredient/brand names is used to build SQL queries to retrieve the drug-event reports in our ADEpedia-on-OHDSI platform. The standardized SQL query for the retrieval of irAE reports is described in Multimedia Appendix 1. In addition, to validate our retrieval query, we also implemented a search by using drug ingredient/brand name verbatim texts and compared the results by different retrieval queries.

For ADE signal detection, the ROR [32] was implemented. ROR is one of the most commonly used disproportionality statistical analysis for signal detection in SRSs such as FAERS [33]. Figure 3 illustrates the contingency table and the equation of the ROR. The ROR value and its 95% CIs were calculated to detect irAE signals. When the case report number was ≥3 and the lower limit of 95% CI of ROR was >1, the signal was considered as a positive irAE signal.

Table 1. The basic information of 6 immune checkpoint inhibitors.

Immune checkpoint inhibitor	Brand name	Food and Drug Administration–approved year	The Observational Health Data Sciences and Informatics concept_id (ingredient/brand name)	RxNorm concept unique identifier (ingredient/brand name)
Ipilimumab	Yervoy	2011	40238188/40238070	1094833/1094837
Pembrolizumab	Keytruda	2014	45775965/45775969	1547545/1547550
Nivolumab	Opdivo	2014	45892628/45892632	1597876/1597881
Atezolizumab	Tecentriq	2016	42629079/42629083	1792776/1792781
Durvalumab	Imfinzi	2017	1594034/1594039	1919503/1919508
Avelumab	Bavencio	2017	1593273/1593278	1875534/1875543

Figure 3. The contingency table and equation for the implementation of the reporting odds ratio.



	Reports with target event	Reports without target event
Reports with immune checkpoint inhibitor	a	b
Reports without immune checkpoint inhibitor	c	d

$$\mathrm{ROR} = \frac{a/b}{c/d}$$

Text-Mining Module

We developed a customized text-mining pipeline to identify irAEs from the text of drug labels and irAEs-related literature using cTAKES v4.0. cTAKES is a widely used clinical information extraction tool that can discover clinically named entities and clinical events using a dictionary lookup algorithm [34]. Moreover, we implemented cTAKES using MedDRA as the dictionary to conduct text mining, so that the adverse events extracted could be standardized by MedDRA-preferred terms (PTs). Note that the result of data mining is *signal* because they have statistical significance, whereas the results extracted by text-mining pipeline are called irAE *terms*.

For drug label text mining, we collected the drug labels of the 6 FDA-approved immune checkpoint inhibitors from the DailyMed website [27]. These drug labels were downloaded in the SPL format, which is a document markup standard approved by HL7 and adopted by the FDA as a mechanism for exchanging product and facility information. Multimedia Appendix 2 shows the drug label links in DailyMed. Then, we extracted the text under the section WARNINGS AND PRECAUTIONS and the section ADVERSE REACTIONS from the SPL files of 6 labels as the dataset of drug label text mining. In addition, to evaluate the performance of cTAKES in our irAE text-mining pipeline, 2 authors (KR and GJ) manually reviewed the text under those 2 sections of the 6 drug labels and identified the irAE terms out of the drug label text, and they reached consensus via discussions. Both KR and GJ have medical backgrounds, and KR is a medical oncologist with both clinical and research expertise in treatment toxicities. The irAE terms identified from the manual review were used as a gold standard to assess the baseline performance of our text-mining pipeline, and standard measures (precision, recall, and *F*-measure) were calculated for the performance evaluation.

To identify irAEs from related literature, we searched the PubMed using the query “immune-related[All Fields] AND adverse[All Fields] AND events[All Fields].” A total of 679 irAE-related studies were found, and the abstracts were downloaded for all search results (as of January 2018). We also extracted the text from the full text of 20 review papers from the search results for text mining. The distribution of irAE-related literature by year is illustrated in Table 2, showing a trend that the number of studies on irAEs has increased significantly in recent years. We then implemented the text-mining pipeline with MedDRA as a dictionary to extract the irAE terms from both the abstracts and the full review papers of the irAE-related literature.

Table 2. The distribution of literature on immune-related adverse events by year (PubMed retrieve date: January 24, 2018).

Publication year	Publication number
2006	1
2007	1
2008	5
2009	7
2010	7
2011	11
2012	11
2013	37
2014	47
2015	74
2016	150
2017	260
2018	68

Signal Filtration Module

We reviewed all irAE signals that were identified from the signal detection and classified them into 3 categories: labeled signals (ie, those signals that could be validated by drug labels), unlabeled published signals (ie, signals that could not be found in drug labels, but in published literature), and new signals (ie, signals that could not be found either in drug labels or published literature). Then, 2 oncologists (KR and AM) manually reviewed the new signals category and gave their comments about whether an irAE signal in that category could be seen as a potentially new signal. Note that those oncologists only reviewed the detection results, and they did not have access to any other clinical data to help them ascertain what might be due to the cancer or the treatment.

Results

Data Standardization Results

After the ETL process, raw FAERS data were loaded into 8 OHDSI CDM tables. A total of 4,619,362 adverse event case reports were transferred into the OHDSI CDM. Among these patients, 2,577,989 (55.81%) were female, 1,603,982 (34.72%) were male, and the sex of 437,391 (9.47%) was unknown/not specified.

Table 3 shows the total numbers of irAE reports in the raw FAERS and CDM-based FAERS. It should be noted that one patient may receive more than one immune checkpoint inhibitor. We found that more irAE reports were collected after the ETL process. In CDM-based FAERS, a total of 24,595 immune checkpoint inhibitor-related AE reports were collected, compared with 24,500 in the raw FAERS before the ETL process. Of the 6 immune checkpoint inhibitors, nivolumab (Opdivo) had the most AE reports (n=12,557 before ETL and n=12,569 after ETL), followed by ipilimumab (Yervoy; n=8264 before ETL and n=8268 after ETL).

Table 3. Total report numbers of 6 immune checkpoint inhibitors.

Immune checkpoint inhibitor	Brand name	Adverse drug event report number (before extract, transform, and load)	Adverse drug event report number (after extract, transform, and load)
Ipilimumab	Yervoy	8264	8268
Pembrolizumab	Keytruda	5020	5099
Nivolumab	Opdivo	12,557	12,569
Atezolizumab	Tecentriq	891	893
Durvalumab	Imfinzi	27	27
Avelumab	Bavencio	5	5
Total reports	N/A[a]	24,500	24,595

[a] N/A: not applicable.

Immune-Related Adverse Events Signal Detection Results

To provide a comprehensive perspective for irAEs, we conducted irAE signal detection at 2 different MedDRA adverse event levels: the system organ class (SOC) level and the PT level. SOC level is the highest level of MedDRA, which contains 27 groupings by etiology (eg, SOC infections and infestations), manifestation site (eg, SOC gastrointestinal disorders), and purpose (eg, SOC surgical and medical procedures). A PT term is a distinct descriptor (single medical concept) that is linked to at least one SOC. Table 4 shows the 7 positive signals detected in the SOC level.

Moreover, 94 positive signals in the PT level were detected in patients who used 1 of the 6 immune checkpoint inhibitor drugs. Among all the positive irAE signals, hypophysitis had the highest ROR value (ROR 5398.8; 95% CI 3105.1-9386.9), followed by hypopituitarism (ROR 135.1; 95% CI 106.7-171.1), blood corticotrophin decreased (ROR 59.5; 95% CI 3105.1-9386.9), adrenal insufficiency (ROR 36.1; 95% CI 31.5-41.3), and colitis (ROR 32.7; 95% CI 30.5-35.0), which means these irAEs were possibly suffered most often by the patients who were immune checkpoint inhibitors.

We also classified the irAE signals using the MedDRA SOCs to obtain a high-level understanding of the distribution of the irAE signals (shown in Table 5). Note that 1 PT might be linked to more than 1 SOC, so the total signal number in Table 5 was more than 94. All the signals we detected at the PT level could be classified into 19 SOCs. Moreover, 14 PT-level signals were categorized in *Respiratory, thoracic and mediastinal disorders*, which is the SOC with the most signals, followed by *Gastrointestinal disorders*, *Cardiac disorders*, *Infections and infestations*, and *Nervous system disorders*, of which SOCs also had more than 10 PT level signals. In addition, there was at least one PT-level signal in each of the 7 SOCs we previously detected as a positive SOC-level signal, which also validated our detection results at the SOC level. The detailed information of the 94 irAE signals is illustrated in Multimedia Appendix 3.

Table 4. The signal detection results at the system organ class level.

Medical Dictionary for Regulatory Activities code	System organ class	Reporting odds ratio (95% CI)
10014698	Endocrine disorders	2.98 (2.84-3.12)
10019805	Hepatobiliary disorders	2.53 (2.39-2.68)
10027433	Metabolism and nutrition disorders	1.76 (1.69-1.83)
10005329	Blood and lymphatic system disorders	1.56 (1.48-1.64)
10029104	Neoplasms benign, malignant, and unspecified (including cysts and polyps)	1.38 (1.30-1.46)
10038738	Respiratory, thoracic, and mediastinal disorders	1.27 (1.23-1.31)
10017947	Gastrointestinal disorders	1.16 (1.12-1.19)

Table 5. System organ class distribution of preferred term–level signals.

System organ class	Signal number
Respiratory, thoracic, and mediastinal disorders[a]	14
Gastrointestinal disorders[a]	13
Cardiac disorders	10
Infections and infestations	10
Nervous system disorders	10
General disorders and administration site conditions	9
Investigations	9
Immune system disorders	8
Endocrine disorders[a]	5
Hepatobiliary disorders[a]	5
Injury, poisoning, and procedural complications	5
Metabolism and nutrition disorders[a]	5
Skin and subcutaneous tissue disorders	5
Blood and lymphatic system disorders[a]	4
Eye disorders	4
Musculoskeletal and connective tissue disorders	4
Vascular disorders	4
Renal and urinary disorders	3
Neoplasms benign, malignant, and unspecified (including cysts and polyps)[a]	1

[a]Represents the system organ class that was detected as a positive signal in the system organ class level.

Text-Mining Results

As mentioned previously, we utilized cTAKES with MedDRA as a dictionary to identify the irAE terms from the drug label of 6 immune checkpoint inhibitors. A total of 421 and 918 irAE terms were found by text mining of drug labels and irAEs-related literature, respectively.

Regarding drug label text mining, we found that most of the irAE terms identified by cTAKES were in the PT level of MedDRA. However, some of the irAE terms were defined as lowest-level terms (LLTs) in MedDRA. An LLT is a synonym, lexical variant, quasi-synonym, subelement, or an identical to its related PT and could be linked to only one PT. To unify the irAE terms to standard concepts at the same level, we mapped all the LLTs into PTs based on the recommendations of MedDRA and FDA. As a result, 490 irAE terms were extracted from the texts of all the 6 drug labels, comprising 474 PTs, 15 SOC, and 1 high-level term (HLT, a superordinate descriptor for the PTs linked to it). More details of the irAE terms identified from drug labels by our text-mining pipeline are provided in Multimedia Appendix 4.

For the text-mining evaluation, as mentioned in the *Methods* section, the irAE terms manually identified by 2 experts (KR and GJ) from drug labels were seen as the gold standard. Then, irAE terms extracted by the text-mining pipeline were compared with the gold standard to acquire the text-mining performance. As the text-mining pipeline, we also linked all the LLT-level irAE terms to the PT level. Using the expert-based manual review process, we identified a total of 421 distinct irAE terms from drug labels of the 6 immune checkpoint inhibitors, comprising 401 PTs, 10 SOCs, 1 HLT, and 9 terms that could not be mapped with MedDRA concepts. Multimedia Appendix 5 provides the details of the manually identified irAE terms. Table 6 shows the distribution of the irAE terms in different drug labels and the performance of our text-mining pipeline. As illustrated in the table, the overall precision, recall, and *F*-measure of our text-mining pipeline are 79.39%, 92.40%, and 85.40%, respectively, which indicates that our pipeline could provide satisfactory text-mining results and achieved the requirement of our irAE identification task.

For irAE-related literature text mining, by using our text-mining pipeline, a total of 918 unique irAE terms (in PT or higher level) were identified from 679 irAE-related abstracts and 20 irAE-related review papers, in which 306 (33.33%) terms were covered by the irAE terms that were extracted in drug labels, and the remaining 612 (66.67%) terms were not covered by the labeled irAE terms. This indicates that some unlabeled terms can be identified from our text-mining pipeline. Multimedia Appendix 6 provides the results of irAE-related literature text mining.

Table 6. Performance of text-mining pipeline for the identification of immune-related adverse events from drug labels of 6 immune checkpoint inhibitors.

Immune checkpoint inhibitor	Manually identified immune-related adverse events terms	Clinical Text Analysis and Knowledge Extraction System–identified immune-related adverse events terms	True positive	False positive	False negative	Precision (TP/[TP+FP]), %	Recall (TP/[TP+FN]), %	*F*-measure (2PR/[P+R]), %
Ipilimumab	122	138	103	35	19	74.6	84.4	79.2
Pembrolizumab	192	228	179	49	13	78.5	93.2	85.2
Nivolumab	215	262	202	60	13	77.1	93.9	84.7
Atezolizumab	142	157	129	28	13	82.2	90.9	86.3
Durvalumab	179	183	156	27	23	85.3	87.2	86.2
Avelumab	146	176	130	46	16	73.9	89.0	80.8
Total	421	490	389	101	32	79.4	92.4	85.4

Signal Filtration Results

To filter the irAE signals we detected, we compared all 94 irAE signals with the text-mining results and then classified all the signals into 3 categories as per our definition in the *Methods* section. Figure 4 shows the overlap of the irAEs terms identified in 3 different mining tasks. In total, 1135 unique irAE terms were identified by CDM-based FAERS data mining, drug label manual review, and irAE-related literature text mining. Out of 94 positive signals in the PT level detected using CDM-based FAERS, 53 signals (56%) were the labeled signals we identified from drug labels, 10 signals (11%) were the unlabeled published signals identified from the literature, and 31 signals (33%) were potentially new signals that were not covered by drug labels and literature (as shown in Table 7). Multimedia Appendix 7 demonstrates the details of labeled signals, unlabeled published signals, and new signals. For a further manual review, 2 oncologists separately marked 15 and 8 signals that were *possibly new*, after reviewing a total of 31 irAE signals in the new signal category. The kappa coefficient value of the review is 0.48, which showed a *moderate agreement* between the 2 oncologists [35]. Moreover, 7 irAE signals were identified as potentially new signals by both the oncologists (as shown in Table 7).

Figure 4. Venn diagram illustrating the immune-related adverse events terms detected from different sources. CDM, common data model; FAERS, Food and Drug Administration’s adverse event reporting system.

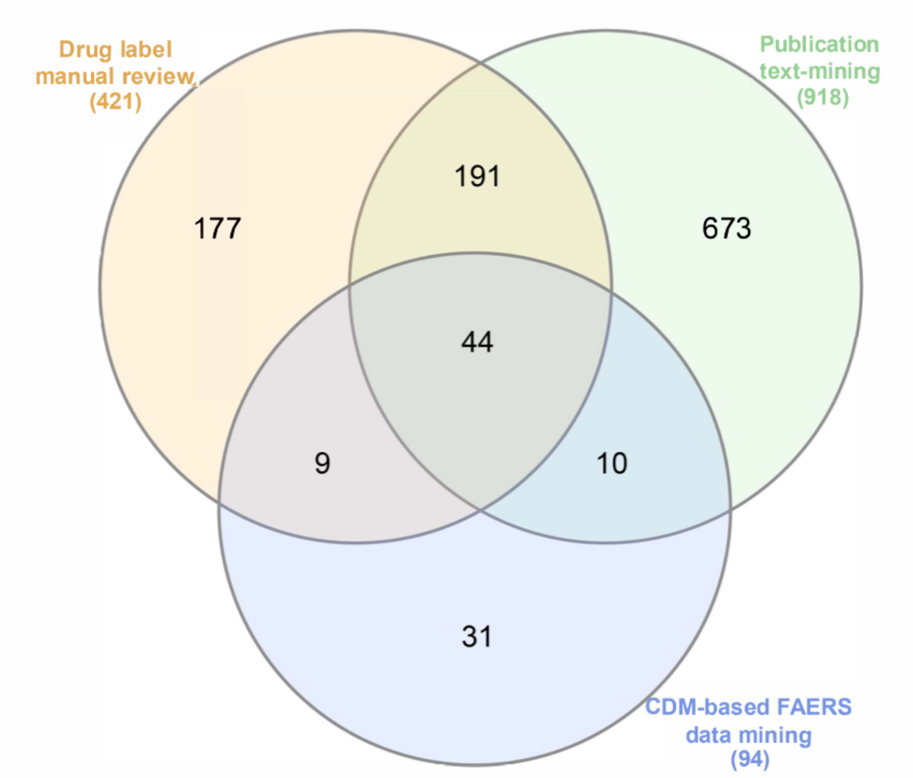


Table 7. A list of 31 potentially new signals not identified in drug labels or literature (ranked by reporting odds ratio).

Medical Dictionary for Regulatory Activities code	Preferred term	System organ class	Reporting odds ratio (95% CI)
10005452	Blood corticotrophin decreased	Investigations	59.49 (34.44-102.74)
10053481	Bronchopleural fistula[a]	Respiratory, thoracic, and mediastinal disorders	19.51 (6.96-54.67)
10006437	Bronchial fistula[a]	Respiratory, thoracic, and mediastinal disorders	19.01 (6.79-53.20)
10042569	Superior vena cava syndrome	Vascular disorders/ neoplasms benign, malignant, and unspecified (including cysts and polyps)	10.62 (5.78-19.51)
10061457	Facial nerve disorder[a]	Nervous system disorders	9.51 (3.48-25.97)
10044291	Tracheal obstruction[a]	Respiratory, thoracic, and mediastinal disorders/ injury, poisoning, and procedural complications	7.83 (2.47-24.87)
10058838	Enterocolitis infectious	Gastrointestinal disorders/infections and infestations	7.64 (3.77-15.51)
10065764	Mucosal infection	General disorders and administration site conditions/infections and infestations	7.13 (2.25-22.59)
10013832	Duodenal perforation	Gastrointestinal disorders	6.50 (2.88-14.68)
10006440	Bronchial obstruction[a]	Respiratory, thoracic, and mediastinal disorders	6.34 (3.13-12.82)
10061145	Eyelid function disorder[a]	Eye disorders	5.73 (1.82-18.09)
10007196	Capillary leak syndrome[a]	General disorders and administration site conditions/vascular disorders	5.62 (2.78-11.35)
10010276	Conduction disorder	Cardiac disorders	4.92 (2.19-11.07)
10036774	Proctitis	Gastrointestinal disorders	4.90 (2.82-8.50)
10021305	Ileal perforation	Gastrointestinal disorders	4.09 (1.30-12.84)
10009995	Colonic fistula	Gastrointestinal disorders	3.81 (1.21-11.95)
10064774	Infusion site extravasation	Injury, poisoning, and procedural complications/general disorders and administration site conditions	3.51 (2.30-5.36)
10051341	Bile duct stenosis	Hepatobiliary disorders	3.45 (1.42-8.35)
10042241	Stridor	Respiratory, thoracic, and mediastinal disorders	3.15 (1.41-7.06)
10035623	Pleuritic pain	Respiratory, thoracic, and mediastinal disorders	3.12 (1.67-5.82)
10025256	Lymphocyte count decreased	Investigations	2.97 (2.29-3.85)
10063057	Cystitis noninfective	Renal and urinary disorders	2.83 (1.05-7.60)
10005630	Blood lactate dehydrogenase increased	Investigations	2.81 (2.10-3.76)
10041549	Spinal cord compression	Nervous system disorders	2.81 (1.66-4.76)
10008612	Cholecystitis	Hepatobiliary disorders	2.59 (1.87-3.58)
10041103	Small intestinal perforation	Gastrointestinal disorders	2.46 (1.02-5.94)
10003662	Atrial flutter	Cardiac disorders	2.45 (1.50-4.02)
10036206	Portal vein thrombosis[a]	Vascular disorders/hepatobiliary disorders	2.43 (1.30-4.53)
10029164	Nephrotic syndrome	Renal and urinary disorders	2.37 (1.42-3.94)
10003673	Atrioventricular block complete	Cardiac disorders	1.85 (1.09-3.13)
10003504	Aspiration	Respiratory, thoracic, and mediastinal disorders	1.60 (1.02-2.51)

[a]Identified as potentially new signals by both oncologist reviewers.

Discussion

Principal Findings

To the best of our knowledge, this is the first comprehensive, novel signal detection and filtration study of irAEs utilizing multiple drug safety data sources. We proposed a framework to detect the irAE signals from a standardized FAERS database and utilized a text-mining pipeline with drug labels and existing literature to discover *potentially new irAE signals*. Our framework could facilitate ADE detection and filtration toward the goal of next-generation pharmacovigilance. This could decrease the labor consumption in new irAE signal selection and provide stronger hypotheses for further experimental validation. In the future, the results of this work will be potentially combined with the EHR data to leverage the real-world discovery of treatment toxicities.

We utilized standard OHDSI CDM to represent the FAERS data (ie, the ADEpedia-on-OHDSI platform) and created standard queries for signal detection, which provides a solid data infrastructure to make the queries portable and signal detection results reproducible. More importantly, through the comparison of data collection between the raw FAERS and CDM-based FAERS, we found that the OHDSI CDM could improve the precision of the data collection. For example, for the drug pembrolizumab, we collected 5099 reports from the CDM-based FAERS, 79 reports more than those we collected from the raw FAERS. To illustrate the reason for the difference in data collection using the OHDSI CDM-based FAERS, we manually checked the data we collected from the raw FAERS and CDM-based FAERS. For example, we discovered that when we utilized the standard OHDSI concept id as a query to retrieve CDM-based FAERS, we could collect more reports regarding the drug name “MK-3475,” which was the original name of pembrolizumab in its early development, in addition to those reports we retrieved when we used the drug ingredient name “Pembrolizumab” and brand name “Keytruda.” This meant that we improved the true positive rate and precision for data collection. Moreover, we could save time for collecting data through a standard query. For example, for pembrolizumab, it took 9.4 seconds to pull all data from CDM-based FAERS with our standard query, in contrast to approximately 70 seconds for the raw FAERS data collection through a fuzzy search query with the drug/brand name terms.

We also leveraged text-mining technology to process unstructured drug safety data. We implemented our text-mining pipeline on the drug label and irAE literature with MedDRA as a dictionary to identify the irAE terms. In addition, to evaluate the performance of our text-mining pipeline, the irAEs in drug labels were manually reviewed and extracted as a gold standard. As a result, the overall precision, recall, and *F*-measure of all 6 drug labels were 79.39%, 92.40%, and 85.40%, respectively. These results indicate that although there were some false positive terms (about 20%) found in our text-mining results, most irAE terms (92.40%) in the text could be extracted correctly by our pipeline. Moreover, we checked the underlying reasons behind the false positive terms. We found that most of these terms are related to the laboratory test name, such as *Alanine aminotransferase*, *Blood alkaline phosphatase*, and so on. Actually, for laboratory tests, the appropriate terms matched with irAEs should be the specific abnormal test result terms, such as *Alanine aminotransferase increased* and *Blood alkaline phosphatase increased*, which were also found in both gold standard and text-mining results. Given this analysis, we plan to improve the precision of our text-mining pipeline using a rule-based approach in a future update.

Limitations and Future Work

Our framework provides an automatic process to detect novel irAE signals that are more valuable for implementing further experimental validation. It also profoundly saves the experts’ time in reviewing drug labels and the literature to filter the known ADEs. In total, we detected 94 irAE signals from FAERS. After the filtering, 31 irAEs were classified into the *new irAE signals* category. In addition, 7 out of 31 signals in the *new signal* category were identified as *potentially new* irAE signals by both the oncologists, which indicated that some of the new signals detected by these algorithms might be false positive. According to the oncologists’ review, some of the signals were marked as *not new*. We consider that one of the main reasons for the false positive signal was that sometimes the description of irAEs by the MedDRA PTs was not so accurate. For example, some detected new signals might be a hyponym of a known irAE, that is, they are more specific than a general irAE. For example, *Conduction disorder* and *Atrioventricular block complete* were detected as new irAE signals by our pipeline. However, the oncologist reviewers judged that these are not new because they are types of arrhythmias, which belong to cardiotoxicity and are known to be associated with the immune checkpoint blockade [36]. Moreover, 2 of the potentially new signals, *bronchopleural fistula/bronchial fistula* and *tracheal obstruction/bronchial obstruction*, are almost the same medical concept. Thus, there is a need to develop a harmonized terminology to report and describe irAEs to interpret safety data more accurately in monitoring missions. One of our previous studies discussed the possibility of leveraging the Common Terminology Criteria for Adverse Events (CTCAE) for irAE standardization. We found that the CTCAE needs an extension to meet the irAE standardization task [37]. Similarly, other studies have also demonstrated how to build a terminology to standardize irAEs [38,39]. In future work, we will improve the text-mining process to facilitate the development of ADE terminology. First, to create a harmonized irAE terminology and make it more suitable for detecting irAE signals from other data sources such as EHR, we will extend the text-mining dictionary to SNOMED CT. Second, we will further improve the performance and automation of our text-mining pipeline to make the terminology easier to update and maintain. Third, we will further evaluate the text-mining pipeline to investigate its feasibility of developing specific terminology sets for other ADE categories.

For those irAE signals in the *new signal* category, some of them were marked as *possibly new* by 2 oncologists because these adverse events may be induced by or associated with cancer, the complication of surgery or radiation, other drugs administered in the cancer treatment regimen, or drug- drug interactions. For example, *Bronchopleural fistula*, *bronchial obstruction*, and *bronchial fistula* all can occur due to a pulmonary cancer or as a complication of pulmonary surgery or radiation [40]. However, FAERS does not provide information such as timeline details about the drug administration/diagnosis/event, which is an obstacle to confirming whether a signal is caused by the treatment or other conditions. Therefore, expert reviews from oncologists are important for our detection pipeline to control false positive signal results. Moreover, as mentioned in the *Methods* section, oncologists also need more clinical data to further validate the relationship between these irAE signals and immunotherapy drugs. Longitudinal observational databases such as EHRs have increasingly been used for further evaluation of adverse event signals. Compared with FAERS data, EHRs not only contain information about patients who suffer ADEs but also provide a more complete medical history of the patients, including treatments, conditions, and potential risk factors. Accordingly, EHRs could be an additional data source for irAE signal detection [41]. We are actively working on integrating EHR data with our ADEpedia-on-OHDSI platform, which can scale to support more advanced signal detection [42]. Our standard-driven platform integrates FAERS data and EHR data together by using the same data standards that could facilitate pharmacovigilance research based on real-world data. Our platform can not only improve data quality but can also facilitate the data collection for comprehensive ADE detection or cross-validation. In the future, we will try to conduct more comprehensive ADE detection studies based on real-world data to overcome the false positive issue. Furthermore, we will also consider utilizing semantic web technology to develop more ADE mining methods.

Conclusions

In this study, we developed and evaluated a novel standards-based framework for signal detection and filtration of irAEs using both the OHDSI CDM and text-mining technologies. We demonstrated that our approach is effective for novel irAE signal detection and filtration; meanwhile, the CDM-based platform provides an infrastructure that would enable the seamless integration of EHR data for improving signal detection in the future.

Acknowledgments

This research was funded in part by National Institutes of Health U01 HG009450. The authors would like to thank Na Hong and Liwei Wang for their previous work in FAERS standardization.

Conflicts of Interest

AM reports research support from Novartis and Verily; remuneration to his institution for participation on advisory boards for AbbVie, Astra Zeneca, BMS, and Genentech; and travel support from Roche and is a nonremunerated director of the Mesothelioma Applied Research Foundation.

Multimedia Appendix 1
The standardized SQL query for the irAE record retrieving.
[DOCX File , 13 KB - medinform_v8i6e17353_app1.docx]

Multimedia Appendix 2
The drug label links of six FDA-approved mAb drugs in DailyMed.
[DOCX File , 13 KB - medinform_v8i6e17353_app2.docx]

Multimedia Appendix 3
The detail information of 94 irAE signals at PT level.
[XLS File (Microsoft Excel File), 44 KB - medinform_v8i6e17353_app3.xls]

Multimedia Appendix 4
The irAE terms identified form drug labels by the text-mining pipeline.
[XLSX File (Microsoft Excel File), 61 KB - medinform_v8i6e17353_app4.xlsx]

Multimedia Appendix 5
The detail of the manually identified irAE terms from the drug labels.
[XLSX File (Microsoft Excel File), 54 KB - medinform_v8i6e17353_app5.xlsx]

Multimedia Appendix 6
The irAE terms identified form irAE-related literature by the text-mining pipeline.
[XLSX File (Microsoft Excel File), 41 KB - medinform_v8i6e17353_app6.xlsx]

Multimedia Appendix 7

The detail of labeled irAE signals, unlabeled published irAE signals, and new irAE signals.

[XLSX File (Microsoft Excel File), 28 KB - medinform_v8i6e17353_app7.xlsx]

References

1. Mellman I, Coukos G, Dranoff G. Cancer immunotherapy comes of age. Nature 2011 Dec 21;480(7378):480-489 [FREE Full text] [doi: 10.1038/nature10673] [Medline: 22193102]
2. Friedman CF, Proverbs-Singh TA, Postow MA. Treatment of the immune-related adverse effects of immune checkpoint inhibitors: a review. JAMA Oncol 2016 Oct 1;2(10):1346-1353. [doi: 10.1001/jamaoncol.2016.1051] [Medline: 27367787]
3. Kaiser J, Couzin-Frankel J. Cancer immunotherapy sweeps nobel for medicine. Science 2018 Oct 5;362(6410):13. [doi: 10.1126/science.362.6410.13] [Medline: 30287641]
4. Postow MA, Sidlow R, Hellmann MD. Immune-related adverse events associated with immune checkpoint blockade. N Engl J Med 2018 Jan 11;378(2):158-168. [doi: 10.1056/NEJMra1703481] [Medline: 29320654]
5. Sharma P, Allison JP. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. Cell 2015 Apr 9;161(2):205-214 [FREE Full text] [doi: 10.1016/j.cell.2015.03.030] [Medline: 25860605]
6. Postow MA, Callahan MK, Wolchok JD. Immune checkpoint blockade in cancer therapy. J Clin Oncol 2015 Jun 10;33(17):1974-1982 [FREE Full text] [doi: 10.1200/JCO.2014.59.4358] [Medline: 25605845]
7. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015 Oct 22;373(17):1627-1639 [FREE Full text] [doi: 10.1056/NEJMoa1507643] [Medline: 26412456]
8. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, KEYNOTE-024 Investigators. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N Engl J Med 2016 Nov 10;375(19):1823-1833. [doi: 10.1056/NEJMoa1606774] [Medline: 27718847]
9. Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. Clin Pharmacol Ther 2012 Jun;91(6):1010-1021 [FREE Full text] [doi: 10.1038/clpt.2012.50] [Medline: 22549283]
10. World Health Organization. The Safety of Medicines in Public Health Programmes - Pharmacovigilance: An Essential Tool. Geneva, Switzerland: WHO Publications; 2006.
11. Garcia CR, Cox JN, Villano JL. Myasthenia gravis and Guillain-Barré syndrome adverse events with immune checkpoint inhibitors. J Clin Oncol 2018 Feb 10;36(5_suppl):37. [doi: 10.1200/jco.2018.36.5_suppl.37]
12. Elias R, Rider J, Tan X, Rahma OE. Single agent and combination checkpoint inhibitors therapy: a post marketing safety analysis. J Clin Oncol 2018 Feb 10;36(5_suppl):125. [doi: 10.1200/jco.2018.36.5_suppl.125]
13. Moslehi JJ, Salem J, Sosman JA, Lebrun-Vignes B, Johnson DB. Increased reporting of fatal immune checkpoint inhibitor-associated myocarditis. Lancet 2018 Mar 10;391(10124):933 [FREE Full text] [doi: 10.1016/S0140-6736(18)30533-6] [Medline: 29536852]
14. Al-Kindi SG, Oliveira GH. Reporting of immune checkpoint inhibitor-associated myocarditis. Lancet 2018 Aug 4;392(10145):382-383. [doi: 10.1016/S0140-6736(18)31542-3] [Medline: 30102167]
15. Oshima Y, Tanimoto T, Yuji K, Tojo A. EGFR-TKI-associated interstitial pneumonitis in nivolumab-treated patients with non-small cell lung cancer. JAMA Oncol 2018 Aug 1;4(8):1112-1115 [FREE Full text] [doi: 10.1001/jamaoncol.2017.4526] [Medline: 29327061]
16. Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf 2009;32(2):99-110. [doi: 10.2165/00002018-200932020-00003] [Medline: 19236117]
17. Council for International Organizations of Medical Sciences. Practical Aspects of Signal Detection in Pharmacovigilance: Report of CIOMS Working Group VIII. Geneva, Switzerland: WHO Publications; 2010.
18. Coloma PM, Trifirò G, Patadia V, Sturkenboom M. Postmarketing safety surveillance: where does signal detection using electronic healthcare records fit into the big picture? Drug Saf 2013 Mar;36(3):183-197. [doi: 10.1007/s40264-013-0018-x] [Medline: 23377696]
19. Huang J, Zhang X, Tong J, Du J, Duan R, Yang L, et al. Comparing drug safety of hepatitis C therapies using post-market data. BMC Med Inform Decis Mak 2019 Aug 8;19(Suppl 4):147 [FREE Full text] [doi: 10.1186/s12911-019-0860-6] [Medline: 31391106]
20. Rahman MM, Alatawi Y, Cheng N, Qian J, Peissig PL, Berg RL, et al. Methodological considerations for comparison of brand versus generic versus authorized generic adverse event reports in the US food and drug administration adverse event reporting system (FAERS). Clin Drug Investig 2017 Dec;37(12):1143-1152 [FREE Full text] [doi: 10.1007/s40261-017-0574-4] [Medline: 28933038]
21. Singh P, Nayernama A, Jones SC, Kordestani LA, Fedenko K, Prowell T, et al. Fatal neutropenic enterocolitis associated with docetaxel use: a review of cases reported to the United States food and drug administration adverse event reporting system. J Oncol Pharm Pract 2019 Oct 8:- epub ahead of print. [doi: 10.1177/1078155219879494] [Medline: 31594460]

22. Xu R, Wang Q. Large-scale combining signals from both biomedical literature and the FDA adverse event reporting system (FAERS) to improve post-marketing drug safety signal detection. BMC Bioinformatics 2014 Jan 15;15:17 [FREE Full text] [doi: 10.1186/1471-2105-15-17] [Medline: 24428898]
23. Xu R, Wang Q. Automatic signal extraction, prioritizing and filtering approaches in detecting post-marketing cardiovascular events associated with targeted cancer drugs from the FDA adverse event reporting system (FAERS). J Biomed Inform 2014 Mar;47:171-177 [FREE Full text] [doi: 10.1016/j.jbi.2013.10.008] [Medline: 24177320]
24. Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. BMC Med Inform Decis Mak 2014 Mar 24;14:13 [FREE Full text] [doi: 10.1186/1472-6947-14-13] [Medline: 24559132]
25. openFDA. FDA Adverse Event Reporting System URL: https://open.fda.gov/data/faers/ [accessed 2020-05-06]
26. OHDSI: Observational Health Data Sciences and Informatics. OHDSI Common Data Model URL: https://ohdsi.org/ [accessed 2020-05-06]
27. DailyMed. DailyMed: NIH URL: https://dailymed.nlm.nih.gov/dailymed/ [accessed 2020-05-06]
28. PubMed-NCBI. PubMed URL: https://www.ncbi.nlm.nih.gov/pubmed/ [accessed 2020-05-06]
29. Yu Y, Ruddy KJ, Hong N, Tsuji S, Wen A, Shah ND, et al. ADEpedia-on-OHDSI: a next generation pharmacovigilance signal detection platform using the OHDSI common data model. J Biomed Inform 2019 Mar;91:103119 [FREE Full text] [doi: 10.1016/j.jbi.2019.103119] [Medline: 30738946]
30. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data 2016 May 10;3:160026 [FREE Full text] [doi: 10.1038/sdata.2016.26] [Medline: 27193236]
31. Athena: OHDSI. License Agreement URL: http://athena.ohdsi.org [accessed 2020-05-06]
32. van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002;11(1):3-10. [doi: 10.1002/pds.668] [Medline: 11998548]
33. Bate A, Evans SJ. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf 2009 Jun;18(6):427-436. [doi: 10.1002/pds.1742] [Medline: 19358225]
34. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
36. Hassel JC, Heinzerling L, Aberle J, Bähr O, Eigentler TK, Grimm M, et al. Combined immune checkpoint blockade (anti-PD-1/anti-CTLA-4): evaluation and management of adverse drug reactions. Cancer Treat Rev 2017 Jun;57:36-49. [doi: 10.1016/j.ctrv.2017.05.003] [Medline: 28550712]
37. Yu Y, Ruddy KJ, Tsuji S, Hong N, Liu H, Shah N, et al. Coverage evaluation of CTCAE for capturing the immune-related adverse events leveraging text mining technologies. AMIA Jt Summits Transl Sci Proc 2019;2019:771-778 [FREE Full text] [Medline: 31259034]
38. Wang Q, Xu R. Immunotherapy-related adverse events (irAEs): extraction from FDA drug labels and comparative analysis. JAMIA Open 2019 Apr;2(1):173-178 [FREE Full text] [doi: 10.1093/jamiaopen/ooy045] [Medline: 30976759]
39. Zini EM, Lanzola G, Quaglini S, Cornet R. Standardization of immunotherapy adverse events in patient information leaflets and development of an interface terminology for outpatients' monitoring. J Biomed Inform 2018 Jan;77:133-144 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.009] [Medline: 29269275]
40. Lois M, Noppen M. Bronchopleural fistulas: an overview of the problem with special focus on endoscopic management. Chest 2005 Dec;128(6):3955-3965. [doi: 10.1378/chest.128.6.3955] [Medline: 16354867]
41. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013 May 1;20(3):413-419 [FREE Full text] [doi: 10.1136/amiajnl-2012-000930] [Medline: 23118093]
42. Yu Y, Ruddy K, Wen A, Zong N, Shintaro T, Chen J, et al. Integrating electronic health record data into the ADEpedia-on-OHDSI platform for improved signal detection: a case study of immune-related adverse events. AMIA 2020 Informatics Summit 2019:710-719 (forthcoming) [FREE Full text]

Abbreviations

ADE: adverse drug event
AEOLUS: adverse event open learning through universal standardization
CDM: common data model
cTAKES: clinical text analysis and knowledge extraction system
CTCAE: common Terminology Criteria for Adverse Events
EHR: electronic health record

ETL: extract, transform, and load
FAERS: Food and Drug Administration’s Adverse Event Reporting System
FDA: Food and Drug Administration
HL7: Health Level Seven
HLT: high-level term
irAEs: immune-related adverse events
LLTs: lowest-level terms
MedDRA: Medical Dictionary for Regulatory Activities
OHDSI: Observational Health Data Sciences and Informatics
PT: preferred term
ROR: reporting odds ratio
SNOMED CT: systematized nomenclature of medicine-clinical terms
SOC: system organ class
SPL: structured product labeling
SQL: Structured Query Language
SRS: spontaneous reporting system
WHO: World Health Organization

Edited by G Eysenbach; submitted 09.12.19; peer-reviewed by R Richesson, S Hume, K Tingay, CH Lin, B Kijsanayotin; comments to author 27.03.20; revised version received 13.04.20; accepted 15.04.20; published 12.06.20.

Please cite as:
Yu Y, Ruddy K, Mansfield A, Zong N, Wen A, Tsuji S, Huang M, Liu H, Shah N, Jiang G
Detecting and Filtering Immune-Related Adverse Events Signal Based on Text Mining and Observational Health Data Sciences and Informatics Common Data Model: Framework Development Study
JMIR Med Inform 2020;8(6):e17353
URL: http://medinform.jmir.org/2020/6/e17353/
doi:10.2196/17353
PMID:32530430

Original Paper

Identification of High-Order Single-Nucleotide Polymorphism Barcodes in Breast Cancer Using a Hybrid Taguchi-Genetic Algorithm: Case-Control Study

Li-Yeh Chuang[1], PhD; Cheng-San Yang[2], MD, PhD; Huai-Shuo Yang[3], MA; Cheng-Hong Yang[3,4,5], PhD

[1]I-Shou Uneiversity, Kaohsiung City, Taiwan

[2]Ditmanson Medical Foundation Chia-Yi Christian Hospital, Chiayi City, Taiwan

[3]Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung City, Taiwan

[4]Drug Development and Value Creation Research Center, Kaohsiung Medical University, Kaohsiung, Taiwan

[5]College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan

Corresponding Author:
Cheng-Hong Yang, PhD
Department of Electronic Engineering
National Kaohsiung University of Science and Technology
No. 415 Jiangong Road, San-Min District
Kaohsiung City, 82778
Taiwan
Phone: 886 7 381 4526
Email: chyang@nkust.edu.tw

Abstract

Background: Breast cancer has a major disease burden in the female population, and it is a highly genome-associated human disease. However, in genetic studies of complex diseases, modern geneticists face challenges in detecting interactions among loci.

Objective: This study aimed to investigate whether variations of single-nucleotide polymorphisms (SNPs) are associated with histopathological tumor characteristics in breast cancer patients.

Methods: A hybrid Taguchi-genetic algorithm (HTGA) was proposed to identify the high-order SNP barcodes in a breast cancer case-control study. A Taguchi method was used to enhance a genetic algorithm (GA) for identifying high-order SNP barcodes. The Taguchi method was integrated into the GA after the crossover operations in order to optimize the generated offspring systematically for enhancing the GA search ability.

Results: The proposed HTGA effectively converged to a promising region within the problem space and provided excellent SNP barcode identification. Regression analysis was used to validate the association between breast cancer and the identified high-order SNP barcodes. The maximum OR was less than 1 (range 0.870-0.755) for two- to seven-order SNP barcodes.

Conclusions: We systematically evaluated the interaction effects of 26 SNPs within growth factor–related genes for breast carcinogenesis pathways. The HTGA could successfully identify relevant high-order SNP barcodes by evaluating the differences between cases and controls. The validation results showed that the HTGA can provide better fitness values as compared with other methods for the identification of high-order SNP barcodes using breast cancer case-control data sets.

(JMIR Med Inform 2020;8(6):e16886) doi:10.2196/16886

KEYWORDS

genetic algorithm; single-nucleotide polymorphism; breast cancer; case-control study

Introduction

Breast cancer has a major disease burden in the female population, with a growing incidence recently [1,2]. Previously, several interpretations of associations between breast cancer and tumor characteristics [3-5], single-nucleotide polymorphisms (SNPs) [6-8], clinicopathological factors [9], and biomarkers [10] revealed relevant association effects between these factors and the risk of cancer. Previous studies also indicated that genomic variation could contribute to the tumorigenicity process in breast cancer [11-14]. Thus, effective approaches for breast cancer estimation are required.

SNPs are crucial genetic variants in genomic association analyses involving leukemia [15], cancers [16], and other diseases [17-19]. Numerous SNPs cannot be excluded from analyses as no relevant differences between cases and controls can be found through conventional methods. Some SNPs may have relevant associations with other SNPs, and these associations are referred to as SNP barcodes. Consequently, the detection of SNP barcodes is vital for association analyses of diseases and cancers [20-23].

An SNP barcode consists of SNPs, and each SNP includes three genotypes. The large space of suitable SNP barcode combinations complicates the statistical evaluation and identification of relevant SNP barcodes. Evolutionary algorithms have been proposed to facilitate statistical identification of SNP barcodes, and a genetic algorithm (GA) is one of the most frequently used algorithms in genomic studies [24,25]. A GA is an effective approach in the identification of relevant genetic associations for various diseases through the use of more efficient search abilities to enhance population diversity [26]. The crossover and local search operations in a GA can reduce the probability of the same vector being identified between two selected SNPs, and hence, they can improve the search ability of this algorithm.

Breast cancer is a major health issue, and machine learning algorithms are frequently employed to detect the complex genomic associations in breast cancer studies. Although previous machine learning approaches could effectively identify SNP associations in genomic studies, the detection rate of SNP barcodes remains challenging for high-order SNP barcodes. Thus, we proposed a hybrid Taguchi-genetic algorithm (HTGA) for high-order SNP barcode identification in a breast cancer case-control study.

Methods

Genetic Algorithm

A GA is a machine learning algorithm inspired by biological evolutionary processes [27]. The first GA operation is population initialization, in which solutions are produced over the solution space; these initial solutions are designated as parents. In the population, two parents are strategically selected according to some fitness values for crossover operators. Crossover operators generate offspring by combining the chromosomal matter from the two parents. Mutation operations can increase population diversity through localized change, eliminating inferior chromosomes from the population and retaining good offspring. Thus, the good factors within the population can be passed on to the next generation. The aforementioned operations and population replacement are repeated until the stopping criterion is satisfied.

Taguchi Method

The methods proposed by Taguchi et al [28] are based on a statistical experimental design to improve the evaluation and performance of products, process conditions, and parameter settings. Taguchi methods primarily rely on orthogonal arrays (OAs) and the signal-to-noise ratio (SNR). An OA is a fractional factorial matrix that provides a comprehensive analysis of interactions among all design factors. This matrix ensures a proportionate comparison of levels for all factors. A two-level OA can be defined as L_n (2^{n-1}), where n=2^k is the number of experimental runs, k (1) is a positive integer, base 2 represents two levels for each design parameter, and $n-1$ is the number of columns in the OA. "*L*" represents "Latin," because the OA experimental design concept is associated with the Latin square. An example of an OA is shown in Table 1.

SNR (η) is used as the selection quality characteristic in the field of communications engineering; it can be used to optimize the parameters for a target. Taguchi methods can classify the parameter design problem into several categories according the problem. Both smaller-the-better and larger-the-better SNR types are used. Considering the set of characteristics y_1, y_2, …, y_n, in the smaller-the-better case, the SNR can be determined using the following equation:

In the larger-the-better case, the SNR can be determined using the following equation:

The SNR evaluates the robustness of the levels of each design parameter. A high-quality result can be achieved for a particular target by controlling the parameters at a particular level with a high SNR value.

Table 1. A L8(27) orthogonal array.

Experiment number	Factors						
	A	B	C	D	E	F	G
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

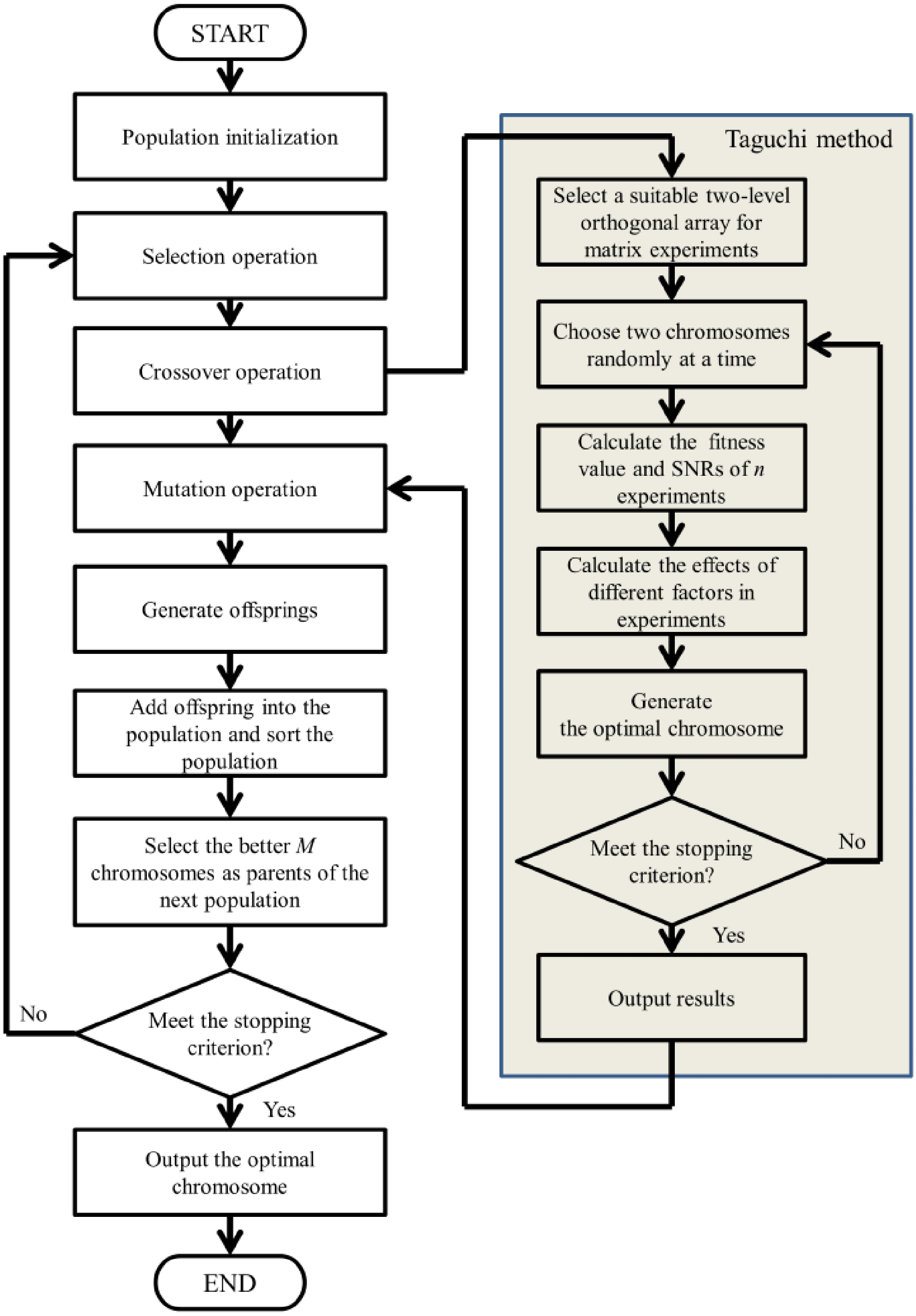
Hybrid Taguchi-Genetic Algorithm

In the HTGA, a Taguchi method is added into GA crossover and mutation operations. Figure 1 depicts a flowchart of the HTGA approach, which includes the below-mentioned 17 steps. The pseudocode of the HTGA is shown in Textbox 1.

HTGA Procedure

The procedure involves the following 17 steps: (1) Population initialization, execute the algorithm and generate an initial population; (2) Fitness value evaluations, evaluate the population’s fitness values; (3) Selection operation, select candidates using the tournament approach; (4) Crossover operation, the probability of crossover is determined by the crossover rate p_c; (5) Select a suitable two-level orthogonal array for the experiment; (6) Randomly choose two chromosomes at a time to execute a matrix experiment; (7) Calculate the function values and SNRs of n experiments in the orthogonal array $L_n\ (2^{n-1})$; (8) Calculate the effects of different factors and in the experiment; (9) An optimal chromosome is generated; (10) Repeat steps 5 through 8 until the expected number $(1/2) \times M \times p_c$ is reached; (11) New chromosomes are generated through the Taguchi method; (12) Mutation operation, mutation probability is determined by the mutation rate p_m; (13) Add chromosomes from a pool into the population; (14) Sort the population by fitness; (15) Select the fittest chromosomes as the new population for the next generation; (16) If the stopping criterion is met, execute step 17; if not, go back to step 2; (17) The chromosome with the highest fitness value is the HTGA solution.

Figure 1. Hybrid Taguchi-genetic algorithm flowchart. SNR: signal-to-noise ratio.


START
Population initialization
Selection operation
Crossover operation
Mutation operation
Generate offsprings
Add offspring into the population and sort the population
Select the better M chromosomes as parents of the next population
Meet the stopping criterion?
No
Yes
Output the optimal chromosome
END
Taguchi method
Select a suitable two-level orthogonal array for matrix experiments
Choose two chromosomes randomly at a time
Calculate the fitness value and SNRs of n experiments
Calculate the effects of different factors in experiments
Generate the optimal chromosome
Meet the stopping criterion?
No
Yes
Output results

Textbox 1. Pseudocode of the hybrid Taguchi-genetic algorithm.

```
Input: maximum iteration as $T$ (termination criterion)
    population size $M$
    crossover rate $P_c$
    mutation rate $P_m$
Output: optimal chromosome (the optimal solution)
begin
# Initialization
/* Initialize $M$ chromosomes as population */
foriteration ← 1 to $T$do
    # Selection operation
    fori ← 1 to $M$do
        /* Randomly select two chromosomes from population as $chromosome_1$ and
        $chromosome_2$ */
        if$fitness_1 \geq fitness_2$do
            winner ← $chromosome_1$
        end if
        else do
            winner ← $chromosome_2$
        end if
            /* put winner into mating pool */
    end for
    # Crossover operation
    fori ← 1 to ($M$ / 2) do
        /* Sequentially select two chromosomes from mating pool as $chromosome_1$
        and $chromosome_2$ */
        if random() < $P_c$do
            crossover($chromosome_1$, $chromosome_2$)
            /* generate two offspring */
        end if
        /* put two offspring into offspring pool */
    end for
    # Taguchi operation
    fori ← 1 to ($0.5 \times M \times P_c$) do
        /* Randomly select two chromosomes from offspring pool as $chromosome_1$
        and $chromosome_2$ */
        Taguchi($chromosome_1$, $chromosome_2$)
        /* generate one offspring */
        /* put one offspring into offspring pool */
     end for
    # Mutation operation
    fori ← 1 to size of offspring pooldo
```

```
            for$j$ ← 1 to dimensions of chromosome do
                if random() < $P_m$do
                    mutation(chromosome[$j$])
                end if
            end for
            /* generate one offspring */
            /* put one offspring into offspring pool */
         end for
        # Replacement operation
        /* Reserve best $M$ chromosomes as new population from population and offspring
        pool */
end for
/* Obtain optimal chromosome */
End
```

Encoding Schemes and Population Initialization

In the proposed GA, a suitable solution to a problem is denoted as chromosome $C = \{c_1, c_2, \ldots, c_n\}$, and the encoding scheme aims to design suitable elements in a chromosome. In the SNP barcode problem, the elements in a chromosome include (1) the indexes of the selected SNPs in the data set and (2) the genotypes of these selected SNPs. Thus, a chromosome C_i is expressed as shown in equation 3.

$$C_i = (SNP_{i,s}, Genotype_{i,g}) \quad \textbf{(3)}$$

where $i = 1, 2, \ldots, m$, and is the population size. $SNP_{i,s}$, where $s = 1, 2, \ldots, n/2$, is a selected SNP dimension in which all SNPs are unrepeatable, and n is the SNP barcode order. $Genotype_{i,g}$ represents the three possible genotypes of the selected $SNP_{i,s}$, where $g = n/2 + 1, n/2 + 2, \ldots, n$ is the selected genotype dimension. In the population initialization, all chromosomes are stochastically generated according to the encoding schemes.

Fitness Function Evaluation

The aim of SNP barcode identification is to detect relevant differences between cases and controls. To optimize the protective effect of the SNP combination, a fitness function is required for comparing cases and controls. A high difference between cases and controls indicates a high probability of detecting relevant SNP barcodes. In the proposed GA, a chromosome is measured by the fitness function shown in equation 4.

$$F(C_i) = number\ (control \cap C_i) - number\ (case \cap C_i) \quad \textbf{(4)}$$

where *number* is the total number of elements in a set, *control* denotes the controls, *case* denotes the cases, and C_i is the *i*th chromosome. Thus, the number of intersections between the *i*th chromosome and the controls is calculated by $number\ (control \cap C_i)$, and the number of intersections between the *i*th chromosome and the cases is calculated by $number\ (case \cap C_i)$. Thereafter, we calculate the difference between $number\ (control \cap C_i)$ and $number\ (case \cap C_i)$ as the fitness value at C_i.

Selection Operation

In the selection operation, a random tournament selection scheme is used to pick each pair of parents from the population [29]. In tournament selection, two chromosomes are randomly selected to compare their individual fitness values. The chromosomes with better fitness values are inserted into the mating pool. According to the mechanism of tournament selection, the probability that the average fitness value of solutions in the mating pool is better than the average fitness value of the parent population is high. Chromosomes in the mating pool are selected for the crossover operation and used to produce offspring. Textbox 2 provides the pseudocode of tournament selection. The selection operation is repeatedly executed until the maximum mating pool size is achieved.

Textbox 2. Tournament selection procedure.

```
Input: population, the list of chromosomes to select from
Input: $chromosome_1$, the first randomly selected chromosome from population
Input: $chromosome_2$, the second randomly selected chromosome from population
Input: $fitness_1$, the fitness value of the first chromosome
Input: $fitness_2$, the fitness value of the second chromosome
Output: winner: the chromosome with better fitness value in tournament
Output: mating pool: reserve the list of chromosomes to execute crossover operation
# Tournament selection
begin
for $i \leftarrow 1$ to size of mating pool do
    Randomly select two chromosomes from population
    if $fitness_1 \geq fitness_2$ do
        $winner \leftarrow chromosome_1$
    end if
    else do
        $winner \leftarrow chromosome_2$
    end else
    put winner into mating pool
end for
```

Crossover Operation

After the selection operation, the crossover operation is implemented to create high-performing individuals. Two chromosomes are sequentially selected from the mating pool as a pair of parents, and then, the crossover operation is executed on them. The crossover operation uses a uniform crossover. Each bit in a chromosome is randomly generated as 0 or 1, and for 1, points are swapped between parent organisms; otherwise, points are not swapped. The encoding schemes establish a single point as an SNP locus with a corresponding genotype locus at the j 2 + 1 position, where $j = 1, 2, \ldots, n/2$ is the index in the chromosome and n is the SNP barcode order. Therefore, $n/2$ bits are randomly generated, and both the j 2 + 1 genotype locus and jth bit representing an SNP are swapped in the parent organisms.

Taguchi Operation

An orthogonal array exhibits Q design factors. Each factor has two levels. An orthogonal array L_n (2^{n-1}) exhibits $n-1$ columns and n individual experiments corresponding to n rows, where $n = 2^k$ and $Q \leq n-1$; k is a positive integer, defined as an integer >1, and it is used for adjusting the number of experimental runs.

The SNR (η) is the mean square deviation of the fitness function. Let two values of η be $\eta_i = (y_i)^2$ and $\eta_i = -(y_i)^2$ (where is negative) in the case of a fitness function that is maximized (larger-the-better). Let y_i be the function evaluation value of experiment $i = 1, 2, 3, \ldots, n$, where n is the number of experiments. The effect of factor f is defined as follows:

E_{fl} = sum of η_i for factor f at level l (**5**)

where i is the experiment number, f is the factor name, and l is the level number.

Mutation Operation

The mutation operation aims to prevent the population from falling into local optima. In all suitable solutions, each offspring element has a chance to undergo a mutation operation. Each mutation position with a probability of mutation p_m generates a random number in (0, 1). If the number is less than p_m at the ith element in an offspring specimen, the ith element will be mutated by a randomly generated possible value.

Replacement Operation

The replacement operation uses an individual to replace the weakest individual in the population. After the completion of the aforementioned operations, the offspring are added to the population, and then, all the parents and offspring are ranked based on their fitness values. Subsequently, top p chromosomes in the population size are selected as the new population for the next generation, where p is the population size.

Termination Condition

The HTGA operation is repeated in successive iterations until the stopping criterion is met. In this study, a maximum number of iterations was used to terminate HTGA operations.

Parameter Setting

This study compared the search effectiveness of the HTGA with that of standard GA, particle swarm optimization (PSO) [30], and chaotic PSO (CPSO) [31] methods. PSO is a swarm intelligence algorithm that simulates the social behavior of organisms. In PSO, each individual represents a particle and considers a potential solution in the swarm population. In CPSO, chaotic theory is incorporated into PSO to increase the search space and enhance PSO performance. PSO and CPSO parameters include population size, iteration size, minimum and maximum inertial weights, and learning factors. In each method, the number of iterations was set to 1000, and the population size was 50 for the test data set. In PSO and CPSO, the minimum and maximum inertial weights were 0.4 and 0.9, respectively. Both weights of learning factors c_1 and c_2 were set to 2. In the tested GA and the proposed HTGA, the probability of crossover (p_c) with an exchange probability was 0.3 and the probability of mutation (p_m) with an exchange probability was 0.05.

Statistical Analysis

The OR was used to evaluate the risk of an SNP barcode [32], and it was defined as follows:

OR = $(TP \times TN) / (FP \times FN)$ (**6**)

where *TP* represents the number of true positives, *TN* represents the number of true negatives, *FN* represents the number of false negatives, and *FP* represents the number of false positives.

Results

Data Sets

A set of 26 SNPs related to growth factor genes was selected to simulate a data set. Several growth factor–related breast cancer genes (*EGF*, *IGF1*, *IGF1R*, *IGF2*, *IGFBP3*, *IL10*, *TGFB1*, and *VEGF*), including 26 SNPs, were used as simulation data to evaluate existing algorithms and the proposed HTGA. The data set only provided the genotype frequencies of each SNP without the original raw data of genotypes. Table 2 presents the SNPs and genotype distributions. The simulated frequencies of SNPs were acquired from the literature [33]. SNPs used in the original data comprised different numbers of individuals; therefore, the number of every SNP must be normalized to the same number. The new data were randomly generated according to the frequency of the original data. All SNP data from the data source were adjusted to 5000 samples for all genotype distributions. For example, for *SNP*1 (gene, *EGF*; dBSNP ref. rs2237054), the total number of three genotypes (ie, TT, TA, and AA) in the control was 2273 (2008 + 259 + 6). The percentage for each genotype in *SNP*1 was calculated as "original data*/sum (%)" (ie, 2008/2273, 88.3% for TT; 259/2273, 11.4% for TA; and 6/2273, 0.3% for AA), where the symbol "*" indicates that the original data were derived from the SNP data set before normalization. On the basis of this percentage, the modified data for *SNP*1 were calculated by multiplying the percentage with the sum of the complete data set (SNP number adjusted to 5000) (ie, 88.3% × 5000 [n=4418] for AA; 11.4% × 5000 [n=569] for Aa, and 0.3% × 5000 [n=13] for aa). Therefore, the modified data for *SNP*1 were adjusted to a total of 5000 (4418 + 569 + 13 = 5000). Thus, 5000 simulation samples of SNP genotypes were randomly generated by following fixed distribution.

Table 2. Estimated effect from individual single-nucleotide polymorphisms of 26 growth factor–related genes for the occurrence of breast cancer.

SNP[a] (gene)	SNP type	Case number/normal number	OR	95% CI	*P* value
1. rs2237054 (*EGF*)	1-TT	4408/4418	N/A[b]	N/A	N/A
1. rs2237054 (*EGF*)	2-TA	570/569	1	0.89-1.14	.97
1. rs2237054 (*EGF*)	3-AA	22/13	1.7	0.85-3.37	.18
2. rs5742678 (*IGF1*)	1-CC	2797/2866	N/A	N/A	N/A
2. rs5742678 (*IGF1*)	2-CG	1844/1837	1.03	0.95-1.12	.52
2. rs5742678 (*IGF1*)	3-GG	359/297	1.24	1.05-1.46	.01
3. rs1549593 (*IGF1*)	1-CC	2924/2970	N/A	N/A	N/A
3. rs1549593 (*IGF1*)	2-CA	1753/1771	1.01	0.93-1.09	.92
3. rs1549593 (*IGF1*)	3-AA	323/259	1.27	1.07-1.50	.008
4. rs6220 (*IGF1*)	1-AA	2643/2698	N/A	N/A	N/A
4. rs6220 (*IGF1*)	2-AG	1933/1951	1.01	0.93-1.10	.80
4. rs6220 (*IGF1*)	3-GG	424/351	1.23	1.06-1.44	.007
5. rs2946834 (*IGF1*)	1-CC	2295/2336	N/A	N/A	N/A
5. rs2946834 (*IGF1*)	2-CT	2171/2150	1.03	0.95-1.12	.53
5. rs2946834 (*IGF1*)	3-TT	534/514	1.06	0.93-1.21	.43
6. rs1568502 (*IGF1R*)	1-AA	2914/2955	N/A	N/A	N/A
6. rs1568502 (*IGF1R*)	2-AG	1840/1807	1.03	0.95-1.12	.46
6. rs1568502 (*IGF1R*)	3-GG	246/238	1.05	0.87-1.26	.65
7. IGF1R-10 (*IGF1R*)	1-AA	3169/3201	N/A	N/A	N/A
7. IGF1R-10 (*IGF1R*)	2-Aa	1545/1582	0.99	0.91-1.08	.77
7. IGF1R-10 (*IGF1R*)	3-aa	286/217	1.33	1.11-1.60	.003
8. rs2229765 (*IGF1R*)	1-GG	1523/1429	N/A	N/A	N/A
8. rs2229765 (*IGF1R*)	2-GA	2533/2489	0.96	0.87-1.05	.33
8. rs2229765 (*IGF1R*)	3-AA	944/1082	0.82	0.73-0.92[c]	.001
9. rs8030950 (*IGF1R*)	1-CC	2737/2745	N/A	N/A	N/A
9. rs8030950 (*IGF1R*)	2-CA	1902/1917	1	0.92-1.08	.92
9. rs8030950 (*IGF1R*)	3-AA	361/338	1.07	0.92-1.25	.41
10. rs680 (*IGF2*)	1-GG	2538/2451	N/A	N/A	N/A
10. rs680 (*IGF2*)	2-GA	2074/2183	0.92	0.85-1.00	.04
10. rs680 (*IGF2*)	3-AA	388/366	1.02	0.88-1.19	.79
11. rs3741211 (*IGF2*)	1-TT	1936/1971	N/A	N/A	N/A
11. rs3741211 (*IGF2*)	2-TC	2367/2269	1.06	0.98-1.16	.17
11. rs3741211 (*IGF2*)	3-CC	697/760	0.93	0.83-1.05	.28
12. IGF2-05 (*IGF2*)	1-AA	2651/2694	N/A	N/A	N/A
12. IGF2-05 (*IGF2*)	2-Aa	1955/1952	1.02	0.94-1.11	.69
12. IGF2-05 (*IGF2*)	3-aa	394/354	1.13	0.97-1.32	.12
13. IGF2-06 (*IGF2*)	1-AA	2160/2162	N/A	N/A	N/A
13. IGF2-06 (*IGF2*)	2-Aa	2237/2284	0.98	0.90-1.07	.66
13. IGF2-06 (*IGF2*)	3-aa	603/554	1.09	0.96-1.24	.21
14. rs2132571 (*IGFBP3*)	1-GG	2415/2407	N/A	N/A	N/A
14. rs2132571 (*IGFBP3*)	2-GA	2163/2157	1	0.92-1.09	.99

SNP[a] (gene)	SNP type	Case number/normal number	OR	95% CI	*P* value
14. rs2132571 (*IGFBP3*)	3-AA	422/436	0.97	0.83-1.12	.65
15. rs2471551 (*IGFBP3*)	1-GG	3225/3284	N/A	N/A	N/A
15. rs2471551 (*IGFBP3*)	2-GC	1591/1515	1.07	0.98-1.17	.13
15. rs2471551 (*IGFBP3*)	3-CC	184/201	0.93	0.76-1.15	.54
16. rs2854744 (*IGFBP3*)	1-AA	1538/1469	N/A	N/A	N/A
16. rs2854744 (*IGFBP3*)	2-AC	2487/2475	0.96	0.88-1.05	.39
16. rs2854744 (*IGFBP3*)	3-CC	975/1056	0.88	0.79-0.99	.03
17. rs2132572 (*IGFBP3*)	1-GG	2908/3027	N/A	N/A	N/A
17. rs2132572 (*IGFBP3*)	2-GA	1805/1728	1.09	1.00-1.18	.051
17. rs2132572 (*IGFBP3*)	3-AA	287/245	1.22	1.02-1.46	.03
18. rs3024496 (*IL10*)	1-TT	1218/1235	N/A	N/A	N/A
18. rs3024496 (*IL10*)	2-TC	2533/2549	1.01	0.92-1.11	.90
18. rs3024496 (*IL10*)	3-CC	1249/1216	1.04	0.93-1.17	.49
19. rs1800872 (*IL10*)	1-CC	3059/3017	N/A	N/A	N/A
19. rs1800872 (*IL10*)	2-CA	1660/1722	0.95	0.87-1.03	.25
19. rs1800872 (*IL10*)	3-AA	281/261	1.06	0.89-1.27	.53
20. rs1800890 (*IL10*)	1-TT	1703/1701	N/A	N/A	N/A
20. rs1800890 (*IL10*)	2-TA	2455/2508	0.98	0.90-1.07	.63
20. rs1800890 (*IL10*)	3-AA	842/791	1.06	0.95-1.20	.32
21. rs1554286 (*IL10*)	1-CC	3400/3446	N/A	N/A	N/A
21. rs1554286 (*IL10*)	2-CT	1431/1410	1.03	0.94-1.12	.54
21. rs1554286 (*IL10*)	3-TT	169/144	1.19	0.95-1.49	.15
22. rs1800470 (*TGFB1*)	1-TT	1850/1914	N/A	N/A	N/A
22. rs1800470 (*TGFB1*)	2-TC	2372/2399	1.02	0.94-1.11	.62
22. rs1800470 (*TGFB1*)	3-CC	778/687	1.17	1.04-1.32	.01
23. rs699947 (*VEGF*)	1-CC	1236/1273	N/A	N/A	N/A
23. rs699947 (*VEGF*)	2-CA	2511/2463	1.05	0.95-1.16	.33
23. rs699947 (*VEGF*)	3-AA	1253/1264	1.02	0.91-1.14	.73
24. rs1570360 (*VEGF*)	1-GG	2278/2341	N/A	N/A	N/A
24. rs1570360 (*VEGF*)	2-GA	2214/2132	1.07	0.98-1.16	.13
24. rs1570360 (*VEGF*)	3-AA	508/527	0.99	0.87-1.13	.92
25. rs2010963 (*VEGF*)	1-GG	2354/2279	N/A	N/A	N/A
25. rs2010963 (*VEGF*)	2-GC	2133/2157	0.96	0.88-1.04	.31
25. rs2010963 (*VEGF*)	3-CC	513/564	0.88	0.77-1.01	.07
26. rs3025039 (*VEGF*)	1-CC	3744/3741	N/A	N/A	N/A
26. rs3025039 (*VEGF*)	2-CT	1160/1174	0.99	0.90-1.08	.81
26. rs3025039 (*VEGF*)	3-TT	96/85	1.13	0.84-1.52	.47

[a] SNP: single-nucleotide polymorphism.

[b] N/A: not applicable.

Comparison of Cases and Controls in Terms of the Effect of a Single SNP

Table 2 compares patients and normal subjects in terms of effect (OR and 95% CI) at a single SNP for growth factor–related genes. Two SNPs within two genes (rs2229765-AA [*IGF1R*] and rs2854744-CC [*IGFBP3*]) showed significant protection associations (rs2229765-AA: OR 0.82, *P*=.001; rs2854744-CC: OR 0.88, *P*=.03) for breast cancer. The minimum and maximum protection associations exhibited ORs of 0.82 and 0.88, respectively, and the other SNPs showed nonsignificant protection associations for breast cancer.

Comparison Between the Proposed HTGA and Existing Algorithms

We compared PSO [34], CPSO [35], and the GA [24] with the HTGA for 2-SNP to 7-SNP barcodes with protection associations (Table 3). ORs (<1) indicate the impact of the protection association of SNP barcodes for the occurrence of breast cancer. A high difference between cases and controls in the SNP barcodes represents informative protection associations, and *P*<.05 indicates a significant difference for the SNP barcode between cases and controls. The identified 3-SNP to 7-SNP barcodes showed that the HTGA provided values with a greater degree of difference as compared with PSO, CPSO, and the GA, indicating that the HTGA identified relevant SNP barcodes with protection associations more effectively (Table 3). HTGA-identified SNP barcodes showed ORs ranging from 0.755 to 0.870 (*P*=.003) for protection associations with breast cancer. The 2-SNP and 3-SNP barcodes in PSO, CPSO, and the GA showed significant differences between cases and controls (2-SNP: *P*=.003, *P*=.001, and *P*=.03, respectively; 3-SNP: *P*=.04, *P*=.04, and *P*=.002, respectively). The 4-SNP barcodes in CPSO and the GA showed significant differences (*P*=.04 and *P*=004, respectively), and the 5-SNP barcode in the GA also showed a significant difference (*P*=.03). Although CPSO and the GA provided better ORs as compared with the HTGA in all SNP barcodes, the degrees of difference indicated that the SNP barcodes identified by the HTGA were superior to those identified by other methods, and *P* values >.05 indicated that these differences revealed by the models were not significant.

Optimization algorithms have been widely applied to detect relevant high-order SNP barcodes in disease and cancer studies [24,25,34]. Differences between cases and controls are often applied to evaluate the values of SNP barcodes in terms of their fitness function design. As indicated in Table 3, the HTGA effectively identified the relevant protection associations of SNP barcodes for breast cancer. The logistic regression analysis results were strongly validated by the outstanding performance of the HTGA in breast cancer SNP barcode identification. The SNP barcodes detected by the proposed HTGA are simply associations between a barcode and disease, and this type of analysis does not support the inference of causality.

Table 3. Estimation of the best protection single-nucleotide polymorphism barcodes for the occurrence of breast cancer as determined by particle swarm optimization, chaotic particle swarm optimization, the genetic algorithm, and the hybrid Taguchi-genetic algorithm.

Order and method	Combined SNP[a]	SNP genotypes	Control number	Case number	Difference	OR	95% CI	*P* value
2-SNP								
PSO[b]	1,8	1-3	941	816	125	0.841	0.76-0.93	.001
PSO	N/A[c]	Other	4059	4184	N/A	N/A	N/A	N/A
CPSO[d]	1,8	1-3	941	816	125	0.841	0.76-0.93	.001
CPSO	N/A	Other	4059	4184	N/A	N/A	N/A	N/A
GA[e]	1,10	1-2	1926	1823	103	0.916	0.85-0.99	.03
GA	N/A	Other	3074	3177	N/A	N/A	N/A	N/A
HTGA[f]	10,17	2-1	1309	1179	130[g]	0.870	0.79-0.95	.003
HTGA	N/A	Other	3691	3821	N/A	N/A	N/A	N/A
3-SNP								
PSO	8,9,22	3-1-2	269	225	44	0.829	0.69-0.99	.043
PSO	N/A	Other	4731	4775	N/A	N/A	N/A	N/A
CPSO	3,8,9	1-3-1	371	319	52	0.850	0.73-0.99	.04
CPSO	N/A	Other	4629	4681	N/A	N/A	N/A	N/A
GA	1,8,15	1-3-1	624	527	97	0.826	0.73-0.94	.002
GA	N/A	Other	4376	4473	N/A	N/A	N/A	N/A
HTGA	1,10,17	1-2-1	1158	1035	123[g]	0.866	0.79-0.95	.003
HTGA	N/A	Other	3842	3965	N/A	N/A	N/A	N/A
4-SNP								
PSO	4,8,14,22	2-3-1-2	99	76	23	0.764	0.57-1.03	.08
PSO	N/A	Other	4901	4924	N/A	N/A	N/A	N/A
CPSO	10,17,21,23	2-1-1-1	268	223	45	0.824	0.69-0.99	.04
CPSO	N/A	Other	4732	4777	N/A	N/A	N/A	N/A
GA	1,10,17,21	1-2-1-1	795	692	103[g]	0.850	0.76-0.95	.004
GA	N/A	Other	4205	4308	N/A	N/A	N/A	N/A
HTGA	1,10,17,21	1-2-1-1	795	692	103[g]	0.850	0.76-0.95	.004
HTGA	N/A	Other	4205	4308	N/A	N/A	N/A	N/A
5-SNP								
PSO	5,6,8,9,26	1-1-3-2-1	91	75	16	0.821	0.60-1.12	.21
PSO	N/A	Other	4909	4925	N/A	N/A	N/A	N/A
CPSO	2,4,8,11,18	1-2-3-1-2	44	32	12	0.726	0.46-1.15	.17
CPSO	N/A	Other	4956	4968	N/A	N/A	N/A	N/A
GA	1,4,15,17,21	1-1-1-1-1	657	585	72	0.876	0.78-0.99	.03
GA	N/A	Other	4343	4415	N/A	N/A	N/A	N/A
HTGA	1,10,17,21,26	1-2-1-1-1	603	520	83[g]	0.846	0.75-0.96	.009
HTGA	N/A	Other	4397	4480	N/A	N/A	N/A	N/A
6-SNP								
PSO	4,8,15,19,22,24	1-3-2-2-1-3	2	0	2	0.333	0.04-3.20	.34
PSO	N/A	Other	4998	5000	N/A	N/A	N/A	N/A

Order and method	Combined SNP[a]	SNP genotypes	Control number	Case number	Difference	OR	95% CI	*P* value
CPSO	3,4,12,16,20,24	1-1-1-2-2-3	28	21	7	0.749	0.43-1.32	.32
CPSO	N/A	Other	4972	4979	N/A	N/A	N/A	N/A
GA	1,2,4,6,15,18	1-1-1-1-1-2	276	247	29	0.900	0.75-1.06	.19
GA	N/A	Other	4724	4753	N/A	N/A	N/A	N/A
HTGA	1,10,15,17,21,26	1-2-1-1-1-1	394	327	67[g]	0.818	0.70-0.95	.01
HTGA	N/A	Other	4606	4673	N/A	N/A	N/A	N/A
7-SNP								
PSO	5,8,11,13,14,24,25	1-1-3-1-1-2-1	6	3	3	0.500	0.13-2.00	.33
PSO	N/A	Other	4994	4997	N/A	N/A	N/A	N/A
CPSO	10,12,16,17,19,22,26	2-2-2-1-2-2-1	27	20	7	0.740	0.41-1.32	.31
CPSO	N/A	Other	4973	4980	N/A	N/A	N/A	N/A
GA	1,2,6,7,10,14,15	1-1-1-3-2-1-1	38	25	13	0.656	0.40-1.09	.10
GA	N/A	Other	4962	4975	N/A	N/A	N/A	N/A
HTGA	1,10,13,15,17,21,26	1-2-2-1-1-1-1	185	141	44[g]	0.755	0.60-0.94	.01
HTGA	N/A	Other	4815	4859	N/A	N/A	N/A	N/A

[a]SNP: single-nucleotide polymorphism.

[b]PSO: particle swarm optimization.

[c]N/A: not applicable.

[d]CPSO: chaotic particle swarm optimization.

[e]GA: genetic algorithm.

[f]HTGA: hybrid Taguchi-genetic algorithm.

[g]The best results in the *n*-SNP barcodes.

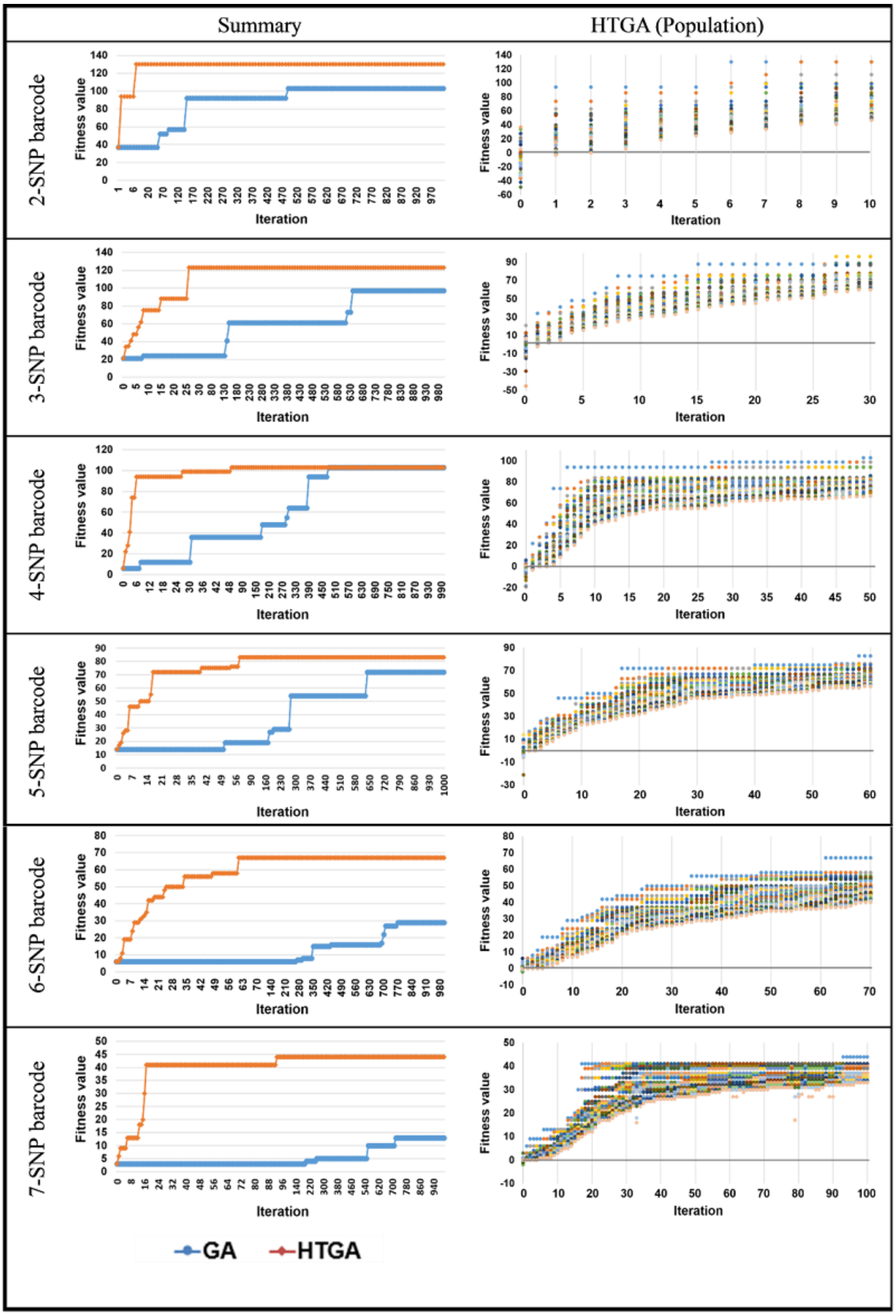
Discussion

Principal Findings

Many breast cancer studies have identified the associations among the effects of important related genes [36-42], including genes related to DNA repair [43,44] and estrogen-response genes [45]. In this study, we introduced a HTGA to identify the SNP barcodes among 26 breast cancer–related SNPs. The HTGA-generated SNP barcodes were examined to determine their protective effects against breast cancer. The results suggest that nonrelevant SNPs might cumulatively reduce the risk of breast cancer, as indicated by the HTGA-generated preventive SNP barcodes. A search space consisting of SNP barcode combinations can generate numerous local optima in multiple regions. These local optima raise challenges for optimization algorithm search operations, because the heuristic and stochastic properties of such optimization algorithms can easily cause searches to become trapped in local optima. A GA population can be updated by referring to other chromosomes to determine the next position in the search space. However, GA operations can result in stagnation if the chromosomes are similar; points of stagnation in a search space are referred to as local optima. The computational processes and comparisons are shown in Figure 2. A Taguchi system is a nonlinear system with deterministic dynamic behavior owing to its ergodic and stochastic properties. Taguchi methods are used to enhance GA crossover operations, and these methods can be remarkably helpful for avoiding population entrapment in local optima because improved solutions can be found through experimentation. Because the population learns from experience, it can be said to exhibit population intelligence. The HTGA can converge quickly to excellent fitness values for SNP barcodes, whereas the GA is very slow to converge and the results are worse than those of the HTGA (Figure 2), indicating that the GA can very easily result in stagnation in regions that may not include any global optima. However, the population is effectively improved in the HTGA, and Figure 2 shows that the fitness values of chromosomes clearly increase over time, proving that the proposed Taguchi method can be used to improve GA performance to identify SNP barcodes. Moreover, our results prove the ability of this Taguchi-based GA to solve SNP barcode identification problems. The optimal parameters of the HTGA could be further analyzed for enhancing the detection ability of SNP barcodes. Our HTGA includes the probability of crossover and mutation. A further investigation with more data sets is required to determine the optimal parameters. Moreover, selection, crossover, mutation, and replacement operations can be analyzed to determine the superior operation strategy for enhancing the ability of our HTGA to detect potential SNP barcodes. If the HTGA is applied for clinical data, we suggest considering permutation testing to examine the relevance of the results obtained. For each trial in permutation testing, the case/control labels would be scrambled, and the algorithm would then search for an optimal solution.

After numerous trials, we would be able to determine the number of times a solution at least as good as the one from the original data is found and then determine if the algorithm is simply fitting the data or identifying underlying associations.

Figure 2. Comparison of improvements to fitness values between the genetic algorithm (GA) and hybrid Taguchi-genetic algorithm (HTGA).The images on the left compare GA and HTGA search results for 1000 iterations. The images on the right present the fitness values of an HTGA population at specific iterations. SNP: single-nucleotide polymorphism.



Conclusions

An HTGA was proposed to effectively identify relevant SNP barcodes among genes related to breast cancer. The study results demonstrated that the HTGA could effectively detect SNP barcodes for problems with numerous high-order SNP barcode combinations. The proposed Taguchi method can improve the GA via the identification of high-dimensional SNP barcodes, and hence, it is integrated following GA crossover operations to systematically optimize chromosomes and thus enhance their values. Moreover, the HTGA can effectively converge to a promising region within the problem space and provide excellent SNP barcode identification. In this study, large data sets were used to evaluate and compare the performances of the GA, PSO, CPSO, and the HTGA, and the results indicated that the HTGA can effectively identify relevant high-order SNP barcodes in breast cancer.

Acknowledgments

This work was supported by the National Science Council, Taiwan (108-2221-E-214-019-MY3 and 108-2221-E-992-031-MY3).

Conflicts of Interest

None declared.

References

1. Sharma R. Breast cancer incidence, mortality and mortality-to-incidence ratio (MIR) are associated with human development, 1990-2016: evidence from Global Burden of Disease Study 2016. Breast Cancer 2019 Jul;26(4):428-445. [doi: 10.1007/s12282-018-00941-4] [Medline: 30604398]
2. Liu F, Lin H, Kuo C, See L, Chiou M, Yu H. Epidemiology and survival outcome of breast cancer in a nationwide study. Oncotarget 2017 Mar 07;8(10):16939-16950 [FREE Full text] [doi: 10.18632/oncotarget.15207] [Medline: 28199975]
3. Abubakar M, Sung H, Bcr D, Guida J, Tang TS, Pfeiffer RM, et al. Breast cancer risk factors, survival and recurrence, and tumor molecular subtype: analysis of 3012 women from an indigenous Asian population. Breast Cancer Res 2018 Sep 18;20(1):114 [FREE Full text] [doi: 10.1186/s13058-018-1033-8] [Medline: 30227867]
4. Visser LL, Elshof LE, Schaapveld M, van de Vijver K, Groen EJ, Almekinders MM, et al. Clinicopathological Risk Factors for an Invasive Breast Cancer Recurrence after Ductal Carcinoma —A Nested Case–Control Study. Clin Cancer Res 2018 Apr 23;24(15):3593-3601. [doi: 10.1158/1078-0432.ccr-18-0201]
5. Park S, Han W, Kim J, Kim MK, Lee E, Yoo T, et al. Risk Factors Associated with Distant Metastasis and Survival Outcomes in Breast Cancer Patients with Locoregional Recurrence. J Breast Cancer 2015 Jun;18(2):160-166 [FREE Full text] [doi: 10.4048/jbc.2015.18.2.160] [Medline: 26155292]
6. Marsaux CF, Celis-Morales C, Livingstone KM, Fallaize R, Kolossa S, Hallmann J, et al. Changes in Physical Activity Following a Genetic-Based Internet-Delivered Personalized Intervention: Randomized Controlled Trial (Food4Me). J Med Internet Res 2016 Feb 05;18(2):e30 [FREE Full text] [doi: 10.2196/jmir.5198] [Medline: 26851191]
7. Shin SJ, You SC, Park YR, Roh J, Kim J, Haam S, et al. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. J Med Internet Res 2019 Mar 26;21(3):e13249 [FREE Full text] [doi: 10.2196/13249] [Medline: 30912749]
8. Yang C, Chuang L, Lin Y. Epistasis Analysis Using an Improved Fuzzy C-Means-Based Entropy Approach. IEEE Trans. Fuzzy Syst 2020 Apr;28(4):718-730. [doi: 10.1109/tfuzz.2019.2914629]
9. Yang C, Moi S, Ou-Yang F, Chuang L, Hou M, Lin Y. Identifying Risk Stratification Associated With a Cancer for Overall Survival by Deep Learning-Based CoxPH. IEEE Access 2019;7:67708-67717. [doi: 10.1109/access.2019.2916586]
10. Moi S, Lee Y, Chuang L, Yuan SF, Ou-Yang F, Hou M, et al. Cumulative receiver operating characteristics for analyzing interaction between tissue visfatin and clinicopathologic factors in breast cancer progression. Cancer Cell Int 2018;18:19 [FREE Full text] [doi: 10.1186/s12935-018-0517-z] [Medline: 29449787]
11. Kotredes KP, Razmpour R, Lutton E, Alfonso-Prieto M, Ramirez SH, Gamero AM. Characterization of cancer-associated IDH2 mutations that differ in tumorigenicity, chemosensitivity and 2-hydroxyglutarate production. Oncotarget 2019 Apr 12;10(28):2675-2692 [FREE Full text] [doi: 10.18632/oncotarget.26848] [Medline: 31105869]
12. Šolman M, Ligabue A, Blaževitš O, Jaiswal A, Zhou Y, Liang H, et al. Specific cancer-associated mutations in the switch III region of Ras increase tumorigenicity by nanocluster augmentation. Elife 2015 Aug 14;4:e08905 [FREE Full text] [doi: 10.7554/eLife.08905] [Medline: 26274561]
13. Derouet M, Wu X, May L, Hoon Yoo B, Sasazuki T, Shirasawa S, et al. Acquisition of anoikis resistance promotes the emergence of oncogenic K-ras mutations in colorectal cancer cells and stimulates their tumorigenicity in vivo. Neoplasia 2007 Jul;9(7):536-545 [FREE Full text] [doi: 10.1593/neo.07217] [Medline: 17710156]
14. Petros JA, Baumann AK, Ruiz-Pesini E, Amin MB, Sun CQ, Hall J, et al. mtDNA mutations increase tumorigenicity in prostate cancer. Proc Natl Acad Sci U S A 2005 Jan 18;102(3):719-724 [FREE Full text] [doi: 10.1073/pnas.0408894102] [Medline: 15647368]
15. Weich N, Ferri C, Moiraghi B, Bengió R, Giere I, Pavlovsky C, et al. GSTM1 and GSTP1, but not GSTT1 genetic polymorphisms are associated with chronic myeloid leukemia risk and treatment response. Cancer Epidemiol 2016 Oct;44:16-21. [doi: 10.1016/j.canep.2016.07.008] [Medline: 27454607]
16. Fu OY, Chang H, Lin Y, Chuang L, Hou M, Yang C. Breast cancer-associated high-order SNP-SNP interaction of CXCL12/CXCR4-related genes by an improved multifactor dimensionality reduction (MDR-ER). Oncol Rep 2016 Sep;36(3):1739-1747. [doi: 10.3892/or.2016.4956] [Medline: 27461876]

17. Tang J, Chuang L, Hsi E, Lin Y, Yang C, Chang H. Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the Apriori algorithm. Biomed Res Int 2013;2013:359634 [FREE Full text] [doi: 10.1155/2013/359634] [Medline: 23984353]
18. Yang C, Wu K, Dahms H, Chuang L, Chang H. Single nucleotide polymorphism barcoding of cytochrome c oxidase I sequences for discriminating 17 species of Columbidae by decision tree algorithm. Ecol Evol 2017 Jul;7(13):4717-4725 [FREE Full text] [doi: 10.1002/ece3.3045] [Medline: 28690801]
19. Yang C, Lin Y, Chuang L. Class Balanced Multifactor Dimensionality Reduction to Detect Gene–Gene Interactions. IEEE/ACM Trans. Comput. Biol. and Bioinf 2020 Jan 1;17(1):71-81. [doi: 10.1109/tcbb.2018.2858776]
20. Ou-Yang F, Lin Y, Chuang L, Chang H, Yang C, Hou M. The Combinational Polymorphisms of ORAI1 Gene Are Associated with Preventive Models of Breast Cancer in the Taiwanese. Biomed Res Int 2015;2015:281263 [FREE Full text] [doi: 10.1155/2015/281263] [Medline: 26380267]
21. Yang P, Ho JW, Yang YH, Zhou BB. Gene-gene interaction filtering with ensemble of filters. BMC Bioinformatics 2011 Feb 15;12(S1). [doi: 10.1186/1471-2105-12-s1-s10]
22. Chuang L, Lane H, Lin Y, Lin M, Yang C, Chang H. Identification of SNP barcode biomarkers for genes associated with facial emotion perception using particle swarm optimization algorithm. Ann Gen Psychiatry 2014;13(1):15. [doi: 10.1186/1744-859x-13-15]
23. Yan R, Cao J, Song C, Chen Y, Wu Z, Wang K, et al. Polymorphisms in lncRNA HOTAIR and susceptibility to breast cancer in a Chinese population. Cancer Epidemiol 2015 Dec;39(6):978-985. [doi: 10.1016/j.canep.2015.10.025] [Medline: 26547792]
24. Chang W, Fang Y, Chang H, Chuang L, Lin Y, Hou M, et al. Identifying association model for single-nucleotide polymorphisms of ORAI1 gene for breast cancer. Cancer Cell Int 2014 Mar 31;14(1):29 [FREE Full text] [doi: 10.1186/1475-2867-14-29] [Medline: 24685237]
25. Chen J, Chuang L, Lin Y, Liou C, Lin T, Lee W, et al. Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. Mitochondrial DNA 2014 Jun;25(3):231-237. [doi: 10.3109/19401736.2013.796513] [Medline: 23777414]
26. Yang C, Moi S, Lin Y, Chuang L. Genetic Algorithm Combined with a Local Search Method for Identifying Susceptibility Genes. Journal of Artificial Intelligence and Soft Computing Research 2016;6(3):203-212. [doi: 10.1515/jaiscr-2016-0015]
27. Holland J. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Cambridge, Massachusetts: MIT Press; 1992.
28. Bendell A, Disney J, Pridmore W. Taguchi Methods: Applications in World Industry. Berlin, Germany: Springer Verlag; 1989.
29. Miller B, Goldberg D. Genetic algorithms, tournament selection, and the effects of noise. Complex Systems 1995;9(3):193-212.
30. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science. 1995 Presented at: Sixth International Symposium on Micro Machine and Human Science; October 4-6, 1995; Nagoya, Japan p. 39-43. [doi: 10.1109/mhs.1995.494215]
31. Chuang L, Chang H, Lin M, Yang C. Chaotic particle swarm optimization for detecting SNP–SNP interactions for CXCL12-related genes in breast cancer prevention. European Journal of Cancer Prevention 2012;21(4):336-342. [doi: 10.1097/cej.0b013e32834e31f6]
32. Mechanic LE, Luke BT, Goodman JE, Chanock SJ, Harris CC. Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. BMC Bioinformatics 2008 Mar 06;9:146 [FREE Full text] [doi: 10.1186/1471-2105-9-146] [Medline: 18325117]
33. Pharoah PD, Tyrer J, Dunning AM, Easton DF, Ponder BA, SEARCH Investigators. Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genet 2007 Mar 16;3(3):e42 [FREE Full text] [doi: 10.1371/journal.pgen.0030042] [Medline: 17367212]
34. Wu S, Chuang L, Lin Y, Ho W, Chiang F, Yang C, et al. Particle swarm optimization algorithm for analyzing SNP-SNP interaction of renin-angiotensin system genes against hypertension. Mol Biol Rep 2013 Jul;40(7):4227-4233. [doi: 10.1007/s11033-013-2504-8] [Medline: 23695493]
35. Chuang L, Chang H, Lin M, Yang C. Chaotic particle swarm optimization for detecting SNP–SNP interactions for CXCL12-related genes in breast cancer prevention. European Journal of Cancer Prevention 2012;21(4):336-342. [doi: 10.1097/cej.0b013e32834e31f6]
36. Chen L, Li W, Zhang L, Wang H, He W, Tai J, et al. Disease gene interaction pathways: a potential framework for how disease genes associate by disease-risk modules. PLoS One 2011;6(9):e24495 [FREE Full text] [doi: 10.1371/journal.pone.0024495] [Medline: 21915342]
37. Yin J, Lu K, Lin J, Wu L, Hildebrandt MA, Chang DW, et al. Genetic variants in TGF-β pathway are associated with ovarian cancer risk. PLoS One 2011;6(9):e25559 [FREE Full text] [doi: 10.1371/journal.pone.0025559] [Medline: 21984931]
38. Ricceri F, Guarrera S, Sacerdote C, Polidoro S, Allione A, Fontana D, et al. ERCC1 haplotypes modify bladder cancer risk: a case-control study. DNA Repair (Amst) 2010 Feb 04;9(2):191-200. [doi: 10.1016/j.dnarep.2009.12.002] [Medline: 20061190]

39. Cauchi S, Meyre D, Durand E, Proença C, Marre M, Hadjadj S, et al. Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value. PLoS One 2008 May 07;3(5):e2031 [FREE Full text] [doi: 10.1371/journal.pone.0002031] [Medline: 18461161]

40. Lin G, Tseng H, Chang C, Chuang L, Liu C, Yang C, et al. SNP combinations in chromosome-wide genes are associated with bone mineral density in Taiwanese women. Chin J Physiol 2008 Feb 29;51(1):32-41. [Medline: 18551993]

41. Yang C, Lin Y, Yen C, Chuang L, Chang H. A systematic gene-gene and gene-environment interaction analysis of DNA repair genes XRCC1, XRCC2, XRCC3, XRCC4, and oral cancer risk. OMICS 2015 Apr;19(4):238-247. [doi: 10.1089/omi.2014.0121] [Medline: 25831063]

42. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, et al. Cumulative Association of Five Genetic Variants with Prostate Cancer. N Engl J Med 2008 Feb 28;358(9):910-919. [doi: 10.1056/nejmoa075819]

43. Conde J, Silva SN, Azevedo AP, Teixeira V, Pina JE, Rueff J, et al. Association of common variants in mismatch repair genes and breast cancer susceptibility: a multigene study. BMC Cancer 2009 Sep 25;9:344 [FREE Full text] [doi: 10.1186/1471-2407-9-344] [Medline: 19781088]

44. Han W, Kim K, Yang S, Noh D, Kang D, Kwack K. SNP-SNP interactions between DNA repair genes were associated with breast cancer risk in a Korean population. Cancer 2012 Feb 01;118(3):594-602 [FREE Full text] [doi: 10.1002/cncr.26220] [Medline: 21751184]

45. Yu J, Hsiung C, Hsu H, Bao B, Chen S, Hsu G, et al. Genetic variation in the genome-wide predicted estrogen response element-related sequences is associated with breast cancer development. Breast Cancer Res 2011 Jan 31;13(1):R13 [FREE Full text] [doi: 10.1186/bcr2821] [Medline: 21281495]

Abbreviations

CPSO: chaotic particle swarm optimization
GA: genetic algorithm
HTGA: hybrid Taguchi-genetic algorithm
OA: orthogonal array
PSO: particle swarm optimization
SNP: single-nucleotide polymorphism
SNR: signal-to-noise ratio

Edited by C Lovis; submitted 04.11.19; peer-reviewed by BTL Brian T Luke, HW Chang; comments to author 15.12.19; revised version received 09.02.20; accepted 08.04.20; published 17.06.20.

Please cite as:
Chuang LY, Yang CS, Yang HS, Yang CH
Identification of High-Order Single-Nucleotide Polymorphism Barcodes in Breast Cancer Using a Hybrid Taguchi-Genetic Algorithm: Case-Control Study
JMIR Med Inform 2020;8(6):e16886
URL: https://medinform.jmir.org/2020/6/e16886
doi:10.2196/16886
PMID:32554381

Original Paper

Understanding Drug Repurposing From the Perspective of Biomedical Entities and Their Evolution: Bibliographic Research Using Aspirin

Xin Li[1,2], PhD; Justin F Rousseau[3], MD, MMSc; Ying Ding[4], PhD; Min Song[5], PhD; Wei Lu[1], PhD

[1]Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan, China

[2]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, United States

[3]Department of Population Health and Department of Neurology, Dell Medical School, The University of Texas at Austin, Austin, TX, United States

[4]School of Information, Dell Medical School, The University of Texas Austin, Austin, TX, United States

[5]Department of Library and Information Science, Yonsei University, Seoul, Republic of Korea

Corresponding Author:
Wei Lu, PhD
Information Retrieval and Knowledge Mining Laboratory
School of Information Management
Wuhan University
299 Bayi DR, Wuchang District
Wuhan, 430072
China
Phone: 86 02768752757
Email: weilu@whu.edu.cn

Abstract

Background: Drug development is still a costly and time-consuming process with a low rate of success. Drug repurposing (DR) has attracted significant attention because of its significant advantages over traditional approaches in terms of development time, cost, and safety. Entitymetrics, defined as bibliometric indicators based on biomedical entities (eg, diseases, drugs, and genes) studied in the biomedical literature, make it possible for researchers to measure knowledge evolution and the transfer of drug research.

Objective: The purpose of this study was to understand DR from the perspective of biomedical entities (diseases, drugs, and genes) and their evolution.

Methods: In the work reported in this paper, we extended the bibliometric indicators of biomedical entities mentioned in PubMed to detect potential patterns of biomedical entities in various phases of drug research and investigate the factors driving DR. We used aspirin (acetylsalicylic acid) as the subject of the study since it can be repurposed for many applications. We propose 4 easy, transparent measures based on entitymetrics to investigate DR for aspirin: Popularity Index (P_1), Promising Index (P_2), Prestige Index (P_3), and Collaboration Index (CI).

Results: We found that the maxima of P_1, P_3, and CI are closely associated with the different repurposing phases of aspirin. These metrics enabled us to observe the way in which biomedical entities interacted with the drug during the various phases of DR and to analyze the potential driving factors for DR at the entity level. P_1 and CI were indicative of the dynamic trends of a specific biomedical entity over a long time period, while P_2 was more sensitive to immediate changes. P_3 reflected the early signs of the practical value of biomedical entities and could be valuable for tracking the research frontiers of a drug.

Conclusions: In-depth studies of side effects and mechanisms, fierce market competition, and advanced life science technologies are driving factors for DR. This study showcases the way in which researchers can examine the evolution of DR using entitymetrics, an approach that can be valuable for enhancing decision making in the field of drug discovery and development.

(JMIR Med Inform 2020;8(6):e16739) doi:10.2196/16739

KEYWORDS

drug repurposing; biomedical entities; entitymetrics; bibliometrics; aspirin; acetylsalicylic acid

Introduction

Background

Despite recent advances in life sciences and technology, drug development is still a costly and time-consuming process with a low rate of success [1]. Discovering a new drug usually takes more than 10 years and costs around $2 billion on average [2]. The number of targetable human genes is approximately 3000, and the identification of serious and even deadly drug side effects is ongoing [3,4]. To overcome these difficulties, many researchers have turned to drug repurposing, which is the practice of identifying novel clinical indicators for existing marketed drugs [5-7].

The past few decades have produced a few successful cases of drug repurposing. For example, sildenafil, originally developed to treat cardiovascular disease, was unexpectedly discovered to be effective against erectile dysfunction [8]. Thalidomide, once used for morning sickness, has been repurposed for the treatment of multiple myeloma [9], and metformin, originally a treatment for type 2 diabetes, has been studied for the treatment of depression, aging, obesity, and even cancer [10,11]. Beta blockers, initially indicated for hypertension, and topiramate, originally used as an antiepileptic, are both repurposed for migraineurs [12,13]. Because of its significant advantages over traditional approaches, in terms of development time, cost, and previous clinical studies, drug repurposing has attracted significant attention from pharmaceutical firms, scientists, and governments in recent years [7,14].

Methodologies for drug repurposing and their successful applications have been widely discussed. Chen et al [15] designed a system-based algorithm called the reverse gene expression score based on several large-scale publicly accessible datasets and demonstrated the potency and efficacy of vorinostat, geldanamycin, and gemcitabine for the treatment of liver cancers. Xu et al [16] found that emricasan had an inhibitory effect on the Zika virus by screening more than 6000 compounds. With the rapid development of natural language processing and deep learning techniques, robust solutions have recently been proposed and have demonstrated potential. Researchers have integrated more than 20 different datasets into a knowledge graph to predict potential drug and target pairs [17-19]. Hamilton et al [20] queried drug-gene-drug interactions within a low-dimensional embedding of biomedical knowledge graphs to predict missing or unobserved links for drug repurposing. Chang et al [21] proposed a novel deep learning model called "CDRscan" that can successfully predict the feasibility of drug repurposing and recommend the most effective anticancer agents for an individual patient. Öztürk et al [22] represented drugs and protein sequences using convolutional neural networks to predict the binding affinities of drug-target interactions.

Academic publications are produced at high volume, with around 3000 new articles currently published per day [23]. No researcher nor clinician can read and comprehend all the relevant articles in their domain [24]. The "known" knowledge has turned into "unknown known" knowledge, with hidden information and patterns waiting to be discovered. This growing body of scholarly data opens a new era of exploiting literature and data to enable data-driven discovery [24]. Literature-based discovery, which connects disconnected entities in literature in PubMed, has been successful in identifying several cases of drug repurposing, such as fish oil for Raynaud's syndrome, magnesium for migraine headaches, and proton pump inhibitors for atrial fibrillation [25-27]. Swanson [26] demonstrated that bibliometrics can be a useful approach to knowledge discovery and recommended that his method could be extended to other disconnected sets of scientific literature to enable cross-disciplinary innovation [28]. With entitymetrics — bibliometric indicators based on entities studied in the medical literature — researchers without domain knowledge can understand the medical function of a drug [29], identify complex undiscovered biological relationships between drugs and targets [30], and detect implicit gene-gene relationships using literature in PubMed [31]. This research demonstrates the potential of applying bibliometrics to medicine to support data-driven discovery. It represents the next generation of bibliometric studies [32] and already shows great promise [33].

Objectives

In this research, we extended bibliometric indicators for biomedical entities mentioned in the PubMed literature to investigate drug repurposing. We used aspirin (salicylic acid) as the target drug. Aspirin is one of the most well-recognized and well-studied drugs with a history dating back to 1500 BC [34]. It was originally used as an analgesic to treat mild to moderate pain. It has been used clinically for the treatment of at least 10 diseases, including coronary artery disease, cerebrovascular disease, peripheral arterial disease, preeclampsia, diabetes, colorectal cancer, Kawasaki disease, Alzheimer's disease, and arthritis [34,35]. New indications for aspirin are still being reported [36-38]. Aspirin has a remarkably wide range of effects and therefore provides an ideal case with which to study drug repurposing. The work described in this paper primarily aimed to identify patterns in the different repurposing phases of aspirin by analyzing the diseases, drugs, and genes related to aspirin. We propose 4 measures based on entitymetrics to identify the characteristics and patterns of drug repurposing for aspirin: Popularity Index (P_1), Promising Index (P_2), Prestige Index (P_3), and Collaboration Index (CI).

Related Work

Drug Repurposing

Drug repurposing has become a dynamic emerging field of drug discovery and development. According to Baker et al [39], in 2018 nearly two-thirds of 35,000 drugs or compounds described in MEDLINE were investigated as potential treatments for diseases other than those for which they were originally indicated. Nearly 200 drugs have been investigated for repurposing for more than 300 diseases. Many successfully repurposed drugs were discovered accidentally, such as the application of thalidomide for multiple myeloma [9] and sildenafil for erectile dysfunction [8].

Approaches have been proposed for the generation of hypotheses about novel drug-target interactions and have been used to develop promising directions for subsequent validation of drug repurposing. In polypharmacology, researchers have proposed 2 types of hypotheses: (1) two drugs could be indicated for the same condition when they produce a similar gene expression profile, and (2) a disease could be one of the indications for a given drug when it has an opposite gene expression profile to that produced by the drug. The Connectivity Map (CMap; Broad Institute, Cambridge, MA), a database for more than 7000 gene-expression profiles of 1309 compounds, has been widely used in this context in previous work. Using a systematic analysis tool, L1000FWD [40], and CMap, Liu et al [41] found that the anticancer drugs KM-00927 and BRD-K75081836 can be used to inhibit histone deacetylase. Kidnapillai et al [42] used gene expression signature data and CMap to identify 10 drugs, including camptothecin, nimesulide, and rescinnamine, that could be effective against bipolar disorder.

In the field of genetics, association analysis has been extensively applied to the interactions between drug targets and diseases to increase the efficiency of drug repurposing. One of the most successful cases in the field of drug repurposing was based on a genome-wide association study (GWAS) [43]. Using GWAS-driven methods, Sanseau et al [44] concluded that 15.6% of genes are the targets of marketed drugs. They found that GWAS traits can be matched with the indications of drugs and genes involved in pathogenesis have a high probability of being targets for drug repurposing. Based on a strong association between the gene TNFSF11 and Crohn’s disease, the authors inferred, and subsequently confirmed, that dishubzumab, originally developed for the treatment of osteoporosis, can be used against Crohn’s disease [44]. Ferrero and Agarwal [45] combined a CMap-based approach with perturbation of transcriptional profiles and disease data from GWAS for target prioritization and drug repurposing. These researchers pointed out that genetic evidence is important in maximizing the success rate of drug repurposing.

These methods in polypharmacology and genetics usually rely on the high-throughput screening of massive amounts of data related to compounds and targets. As knowledge about drug targets accumulates and computational chemistry rapidly develops, simulations of the interactions between drugs and proteins have shown the potential to replace traditional high-throughput screening. Dakshanamurthy et al [46] proposed a proteochemometric method called “train, match, fit, streamline” to conduct molecular docking of over 3000 FDA-approved compounds across the crystal structures of more than 2000 human targets. They found that mebendazole could be used for the inhibition of VEGFR2 kinase and that celecoxib was a promising therapy for malignancies because it binds an adhesion molecule, cadherin-11. Li et al [47] designed a standalone approach to dock over 30 crystal structures of $MAPK_{14}$ and BIM-8 with all drugs from DrugBank and found that nilotinib, as a potential inhibitor of $MAPK_{14}$, could be a cure for inflammatory diseases.

Another significant source of drug repurposing is drug side effects. Typical instances of side effect–based drug repurposing include the use of sildenafil for erectile dysfunction [8] and the application of exenatide acetate for obesity [48], both of which were “happy accidents.” Recently, Yang and Agarwal [49] generated human phenotypic profiles for drugs based on over 3000 side-effect relationships extracted from PharmGKB and employed naïve Bayes methods to identify new indications for drugs according to their side effects. This study also suggested that the use of side effects is a type of clinical phenotypic assay and side effects should be rationally investigated to predict repurposing opportunities for drugs. Ye et al [50] contend that drugs with similar side effects could share the same indications because they may have the same or similar mechanisms of action. Using a side effect similarity–based drug-drug network, they transformed drug repurposing into an information retrieval issue and successfully obtained the top 5 indications of 1234 drugs approved by the FDA.

With the rise of machine learning and deep learning in computer science and bioinformatics, the problem of drug repurposing has been addressed using approaches such as classification [51,52], link prediction [53,54], entity prediction [53], and path prediction [18,55]. Liang et al [53] represented biomedical entities and their relationships in a heterogeneous network using graph2vec and knowledge2vec [56] and employed a cascade learning model to find potential interactions between drugs, genes, diseases, and treatments. They found that vitamin D could be a treatment for prostate cancer. Fu et al [55] treated drug repurposing as a binary classification problem and combined the metapath-based topological features of biomedical entities in Chem2Bio2RDF and a supervised machine learning model to predict links between drugs and targets. They found that the intrinsic feature selection Random Forest algorithm can be valuable for selecting significant topological features for the prediction of links between drugs and genes.

Big Scholarly Data for Medical Knowledge Discovery

Traditionally, knowledge discovery in medical domains has relied on first-hand observation such as epidemiological statistics, follow-ups, and laboratory-generated experimental data [24]. A large number of research papers are published daily, posing significant challenges for scientists wishing to have a comprehensive understanding of their domain [24]. The “known” knowledge has turned into “undiscovered public knowledge,” with patterns and information waiting to be uncovered. This large body of literature and data also provides rich opportunities for researchers to undertake data-driven knowledge discovery. The usefulness of literature-based discovery has been demonstrated in many previous research projects. For instance, the “ABC” model proposed by Swanson in 1986 [25] was used to discover relationships between biomedical entities, such as Raynaud’s syndrome and fish oil [25], migraine headaches and magnesium [26], and atrial fibrillation and proton pump inhibitors [27]. The “ABC” model is co-occurrence–based and is based upon the premise that seemingly unrelated concepts A and C could be related when there is a concept B related to both A and C [27]. Since Swanson’s research, various modifications of the “ABC” model have been proposed to discover hidden relationships among biomedical concepts in PubMed, such as ontology-based entity mapping [57], network-based entity extraction [58], and semantic path–based storytelling [59]. The “ABC” model and its variants indicate that bibliometrics can be a valuable method for medical knowledge discovery in the era of big scholarly data.

Knowledge graphs of big scholarly data can contain nodes representing biomedical entities such as diseases, drugs, genes, pathways, and cell lines and non-biomedical entities such as authors, institutions, articles, journals, conferences, and keywords. Edges in the graph can represent the relationships between the biomedical entities in the literature. Lv et al [60] established a therapeutic knowledge graph for autism using drug entities and MeSH terms extracted from about 20,000 articles relating to autism published between 1946 and 2015. They proposed a novel topology-driven method incorporating various graph-analytical techniques for drug discovery and concluded that tocilizumab, sulfisoxazole, tacrolimus, and prednisone were promising for the treatment of autism. Ding et al [29] constructed an entity-entity citation graph to highlight the significance of biomedical entities embedded in literature for future knowledge discovery. Researchers have also integrated big scholarly data with other publicly accessible biomedical datasets, such as DrugBank [61], Gene Ontology [62], and SIDER [63], to form a comprehensive knowledge graph for medical knowledge discovery. A typical example is the Chem2Bio2RDF database, created by integrating more than 20 chemogenomic datasets with PubMed. Wang et al [30] proposed a novel algorithm called Bio-LDA to automatically extract latent topics in life sciences and identified relationships and patterns among compounds, genes, and diseases from Chem2Bio2RDF. He et al [64] designed a graph-mining algorithm to predict potential relationships between different biomedical entities. The case they studied demonstrated that the antidiabetic drug rosiglitazone has cardiovascular-related side effects.

Entitymetrics, an entity-driven bibliometric method, and the next generation of citation analysis [29,32] make it possible for researchers without domain knowledge to measure the impact, usage, and transfer of knowledge entities embedded in the academic literature for further knowledge discovery [32]. Ding et al [29] built an entity-entity citation graph based on articles related to metformin and detected most of the known interactions of metformin with biomedical entities. Williams et al [65] recognized and quantified relationships between academic discoveries and major advances in the domain of two new drugs, ipilimumab and ivacaftor, to enhance government support and public understanding. Zhu et al [66] established paper-entity, entity-entity co-occurrence, and entity-specific networks based on the scientific literature to investigate the evolution of hepatic carcinoma at a granular level. Lv et al [60] discovered new indications for drugs using topology-driven trend analysis of drug-drug and drug-indication networks. These studies demonstrate the potential of the application of bibliometric methods to data-driven discovery in medical domains.

Drug repurposing, as one of the most significant issues in the field of medical knowledge discovery, has been extensively investigated [17,23,24,27,28,55-57,64]. In this research, we extended the bibliometric indicators for biomedical entities described in the PubMed literature to understand the process of drug repurposing.

Methods

Data Collection

Papers on aspirin-related research published between 1951 and 2018 were collected from PubMed. Since aspirin is known by many names, the search terms were chosen from DrugBank, RxNorm, and MeSH terms [33,61]. The final search query is shown in Textbox 1. Non-journal articles, non-English articles, letters, and editorial commentaries were excluded. In total, 63,387 publications from PubMed were downloaded in XML format.

To better understand the drug repurposing process of aspirin, the relevant research was divided into 4 phases based on previous studies [34,35] and expert suggestions: (1) 1951-1960, the original use; (2) 1961-1990, in-depth studies of pharmacological mechanisms and side effects; (3) 1991-2000, repurposing for cardiovascular diseases; and (4) 2001-2018, repurposing for other diseases, such as colorectal cancer and breast cancer. These phases can also be observed from the evolution and trends of the publications, as shown in Figure 1 and Table 1.

Before extracting biomedical entities, all articles were parsed to obtain PMIDs, publication years, titles, abstracts, authors, journals, and institutions using a dom4j XML parser written in Java. Then, we used spaCy for preprocessing (such as removing the punctuation and stop words) of titles and abstracts in the natural language processing pipeline. In addition, a novel and reliable method of author name disambiguation proposed by Lerchenmueller and Sorenson [67] was used to count distinct authors.

Textbox 1. Search query used for retrieving aspirin-related publications.

(((aspirin) OR (acetylsalicylic acid) OR (acid, acetylsalicylic) OR (“2-(acetyloxy)benzoic aci”) OR (acylpyrin) OR (aloxiprimum) OR (colfarit) OR (dispril) OR (easprin) OR (ecotrin) OR (endosprin) OR (magnecyl) OR (micristin) OR (polopiri) OR (polopiryna) OR (solprin) OR (solupsan) OR (zorprin) OR (acetysal) OR (2-acetoxybenzenecarboxylic acid) OR (2-acetoxybenzoic acid) OR (acetylsalicylate) OR (acetylsalicylsäure) OR (“acide 2-(acétyloxy)benzoïqu”) OR (acide acétylsalicylique) OR (ácido acetilsalicílico) OR (acidum acetylsalicylicum) OR (aspirina) OR (azetylsalizylsäure) OR (o-acetoxybenzoic acid) OR (o-acetylsalicylic acid) OR (o-carboxyphenyl acetate) OR (salicylic acid acetate)) AND (“1951”[PDAT] : “2018”[PDAT]))

Figure 1. Number of aspirin-related studies in PubMed over time. The background colors indicate the 4 phases of aspirin research.

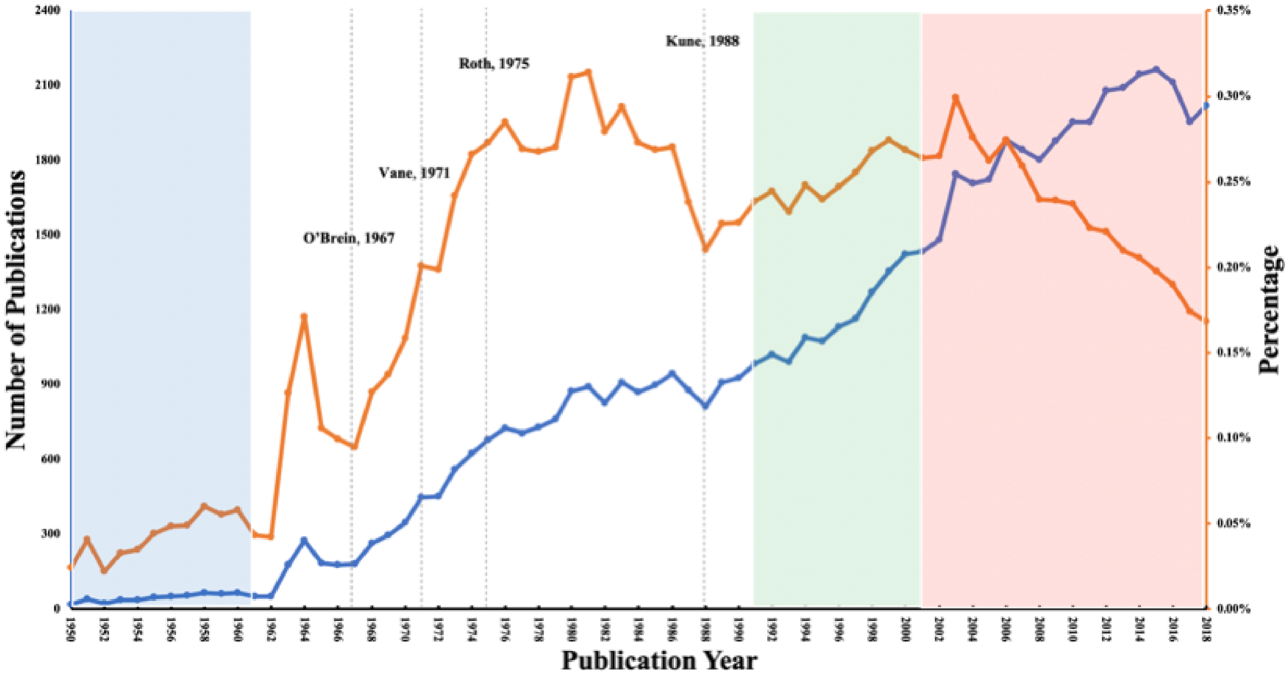


Table 1. Descriptive statistics of the 4 phases of aspirin research.

Phases, Time span	Number of publications	Number of authors	Average number of authors	Number of journals
1. Original use				
1951-1955	208	318	1.76	117
1956-1960	299	498	1.88	159
1951-1960	507	794	1.83	218
2. In-depth studies of pharmacological mechanisms and side effects				
1961-1965	748	1310	2.01	301
1966-1970	1268	2167	2.12	418
1971-1975	2766	4880	2.40	696
1976-1980	3797	7419	2.71	895
1981-1985	4395	10,011	3.16	1033
1986-1990	4470	11,600	3.50	1101
1961-1990	17,444	31,787	2.90	2153
3. Repurposing for cardiovascular diseases				
1991-1995	5164	14,044	3.69	1256
1996-2000	6353	17,694	4.10	1314
1991-2000	11,517	28,818	3.91	1798
4. Repurposing for other diseases				
2001-2005	8099	27,784	4.22	1719
2006-2010	9366	35,313	4.94	1974
2011-2015	10,436	44,603	5.78	2410
2016-2018	6018	30,796	6.73	1881
2001-2018	33,919	118,857	5.33	3865
Total	63,387	171,559	4.39	5443

Biomedical Entity Extraction

The biomedical entity extraction module provided by the biomedical entity search tool (BEST) [68], a dictionary-based biomedical information extraction tool based on sophisticated information retrieval approaches, was deployed to extract entities such as diseases, drugs, and genes. The BEST dictionary is built from 12 different public sources, including NCBI Entrez Gene, DrugBank, T3DB, Animal TFDB, Therapeutic Target DataBase, PubChem, and MeSH [68]. We obtained 1472 unique disease names, 1640 unique drug names, and 3184 unique gene names from the titles and abstracts. Table 2 shows the top 10 biomedical entities of 3 different types and their frequency of appearance in PubMed articles.

Table 2. Top 10 biomedical entities in aspirin-related publications during 1951-2018.

Rank	Diseases	Frequency of diseases	Drugs	Frequency of drugs	Genes	Frequency of genes
1	Coronary disease	2707	Clopidogrel	6223	COX-2	3957
2	Asthma	2277	Ticlopidine	5433	CD143	1495
3	Diabetes	1840	Heparin	4391	COX-1	1179
4	Hypersensitivities, drug	1342	Indomethacin	3462	Plasminogen	1131
5	Ulcer, gastric	1146	Warfarin	3457	LDLCQ3	1081
6	Cerebral ischemia	1135	Vitamin F	2760	LPLA2	1047
7	Intracranial vascular disorder	1133	Dipyridamole	2232	GPIIb	1017
8	Ischemic heart disease	1090	Adenosine	2188	$P2Y_{12}$	855
9	Carcinomas, colorectal	1085	Acetaminophen	2099	tPA	748
10	Rheumatoid arthritis	832	Prostacyclin	1498	TNF-α	629

Entitymetric Indicators for Biomedical Entities (P3C)

In order to quantify and visualize the academic importance of individual biomedical entities, 4 transparent and easy entitymetric indexes (P3C) were developed: Popularity Index (P_1), Promising Index (P_2), Prestige Index (P_3), and Collaboration Index (CI). These indicators can be considered as the extensions of the indicators proposed by Kissin and Edwin [33] and Kissin [69] for measuring the academic interest of a drug or technique at the article level. In this study, we adapted the indicators from the perspective of biomedical entities with the goal of understanding drug repurposing. Different from Kissin's indicators, our indicators not only focus on the articles on a given drug but also consider the changes in indicators of biomedical entities (eg, diseases, drugs, and genes) and non-biomedical entities (eg, authors) that are related to the given drug. Detailed explanations of these measures are provided in the following sections.

Popularity Index (P_1)

The P_1 of a certain biomedical entity reflects the percentage of publications discussing that biomedical entity among all publications in a research field during a specific period, usually 5 years. The popularity of a biomedical entity *i*, P_1 (i), is given by:

$$P_1\ (i) = (N_i\ /\ N_T) * 100\%\ \textbf{(1)}$$

where N_i is the number of publications relating to an entity *i* in a period, and N_T represents the total number of publications in the research field during the same period. An increase in P_1 indicates growing academic interest in *i* in the field.

Promising Index (P_2)

The P_2 of a biomedical entity is the change in the popularity of an entity *i* in a research field between two continuous periods. The promising index of a specific biomedical entity *i*, P_2 (i), is expressed as:

$$P_2\ (i) = (N_i\ /\ N_T) - (N_{pi}\ /\ N_{pT})\ \textbf{(2)}$$

where (N_{pi} / N_{pT}) refers to the popularity of the entity *i* in the research field during a previous period of the same length as N_i. P_2 reflects the change in the academic interest in a biomedical entity in a research field in two time periods. When P_2 (i) > 0, it means the academic interest in *i* has increased and vice versa.

Prestige Index (P_3)

P_3 is defined as the ratio of the number of publications about a specific biomedical entity published in the top journals compared to the number of publications about the same entity in all journals that were indexed by PubMed during the same time period. The prestige of a biomedical entity *i*, P_3 (i), is calculated as:

$$P_3\ (i) = (N_{H20}\ /\ N_i) * 100\%\ \textbf{(3)}$$

where N_{H20} represents the number of publications on *i* in the top 20 journals during the same period as N_i. In this study, the top 20 journals were selected based on the journal impact factor and specialty areas. These journals can be divided into two categories: multidisciplinary journals and specialty journals. Fourteen multidisciplinary journals, including JAMA, The Lancet, BMJ, and similar publications, are common for all diseases, drugs, and genes that were studied in this paper. The other 6 journals, such as Circulation, Blood, and The European Heart Journal, are highly associated with aspirin-related specialty areas. The full list of the top 20 journals is shown in Multimedia Appendix 1. P_3 reflects the potential significance of a specific biomedical entity. Continuing high prestige scores could be an early signal of the success of entity-related drug discovery or repurposing [69]. We employed a threshold of 5% to indicate that P_3 was of interest [69].

Collaboration Index (CI)

The CI of a biomedical entity reflects the percentage of the number of distinct authors of articles discussing this entity out of all the distinct authors in the research domain over a period of time. The CI of a biomedical entity *i*, CI (i), is calculated by:

$$\text{CI (i)} = (N_{AI} / N_{AT}) * 100\% \quad \textbf{(4)}$$

where N_{AI} is the number of distinct authors of the publications relating to *i* in a period, and N_{AT} represents the total number of distinct authors in the field in the same period. The CI reflects the research strength of entity *i* in a research field, and a threshold of 5% indicates a level of interest [69].

Results

Overview of Aspirin-Related Studies

Figure 1 shows an overview of aspirin-related research in PubMed from 1951 to 2018. The red and blue lines represent the percentage and absolute numbers, respectively, of articles in PubMed per year. The details of the publications, authors, and journals are shown in Table 1. During the evolution of aspirin, Phase 1 (1951-1960) produced 507 articles, most of which were published in journals covering pharmacy-related or general medicine–related topics (Table 1 and Multimedia Appendix 2). Research in Phase I focused on the anti-inflammatory and antipyretic uses of aspirin, and this phase marks the original use of aspirin.

In Phase 2 (1961-1990), a turning point can be identified in 1967, after which the number of relevant papers per year grew dramatically until 1986. Several significant pharmacological discoveries related to aspirin occurred during this period, including the antiplatelet effect [70], mechanism of inhibition of prostaglandin synthesis [71], and acetylation of the cyclo-oxygenase enzyme [72]. The percentage of aspirin-related articles in PubMed reached its peak in 1981, at about 0.32%, and then decreased. Kune et al [73] reported that aspirin could effectively reduce the incidence of colorectal cancer, after which the percentage began to rise again. After 1975, articles began to occur frequently in journals covering specialty areas, such as Circulation and Thrombosis Research. We identify this phase as the in-depth investigation of the pharmacological mechanisms and side effects of aspirin.

In Phase 3 (1991-2000), there was a steady and stable growth in the number and percentage of aspirin-related articles per year in PubMed (Figure 1). Compared to the first 10 years (1951-1960), there was a >22-fold increase in the number of articles as well as a >36-fold increase in the number of distinct authors. As shown in Multimedia Appendix 2, in both 1991-1995 and 1996-2000, 4 of the top 5 journals were cardiovascular-related journals. We thus identify this phase as repurposing for cardiovascular diseases.

In Phase 4 (2001-2018), the number of articles per year grew continuously and reached its peak (2164) in 2015, but the percentage significantly reduced (Figure 1). From the information presented in Table 1, we note that the numbers of articles, distinct authors, and journals in Phase 4 were all higher than those in the previous 3 periods. The average number of authors in this period had exceeded the total average (4.39). Journals covering other topics, for example Cancer Management and Research, Drugs & Aging, and World Neurosurgery, were increasingly represented (Multimedia Appendix 2), demonstrating that aspirin had been experimentally applied to many other diseases. We thus mark this phase as repurposing for other diseases.

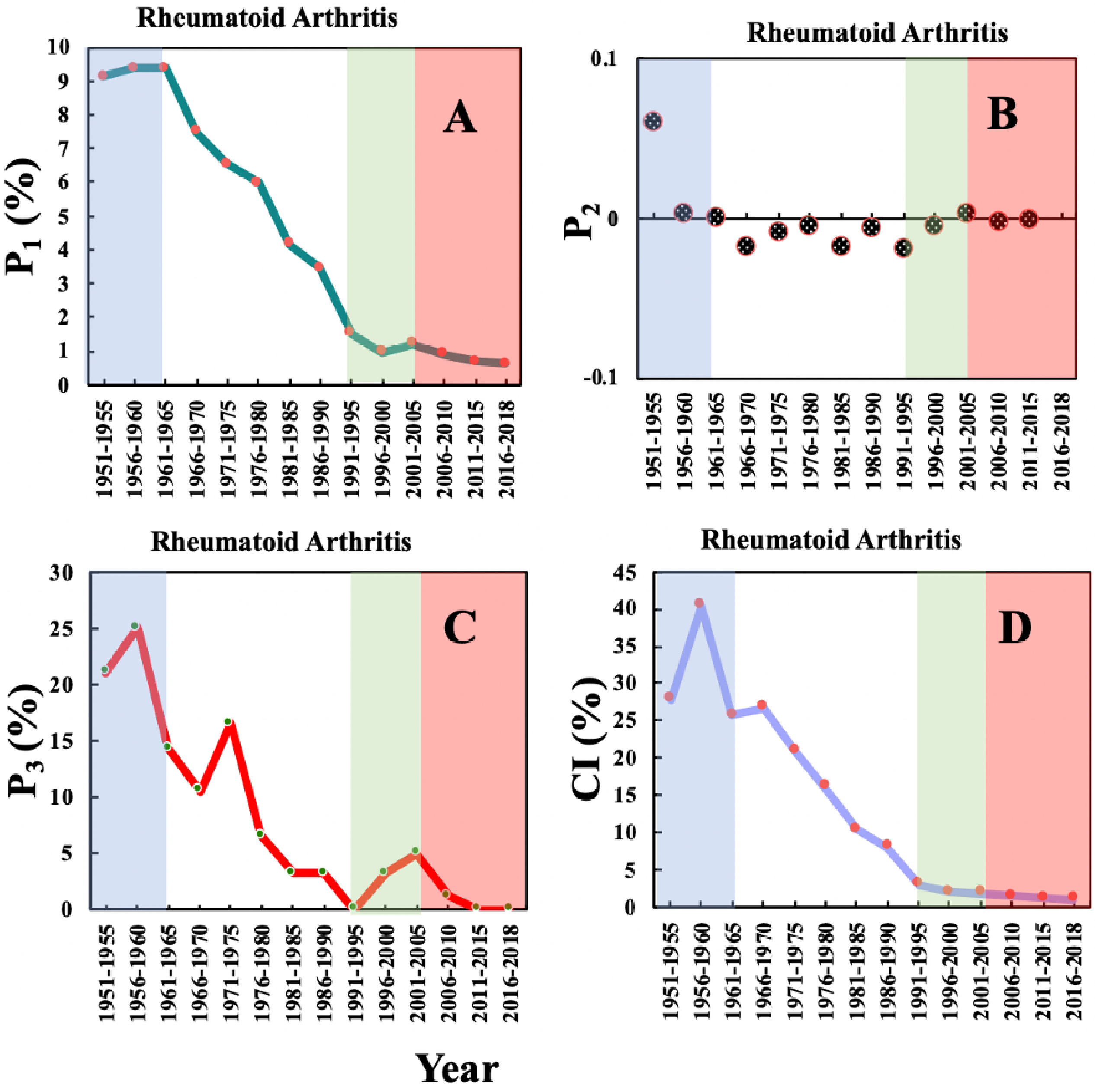
To analyze drug repurposing through all 4 phases from the biomedical entity perspective, we first computed the P3C indicators of the top 10 diseases, drugs, and genes in the cohort of aspirin articles during the period 1951-2018. The results show that there are distinct patterns of these indicators in different repurposing phases. To describe these patterns in detail, we reorganized the 30 biomedical entities (the top 10 diseases, top 10 drugs, and top 10 genes) into the 4 phases of aspirin research, according to when each achieved its maximum P_1, which indicates the focus of research in the field of aspirin. In each phase, we further analyzed the change patterns of the P3C indicators for the most popular biomedical entities, to investigate the features of different phases of drug repurposing, association between entities and P3C indicators, and possible factors driving drug repurposing at the biomedical entity level.

Before Repurposing

Only “rheumatoid arthritis” (RA) reached its maximum P_1 in Phase 1, at 9.36%, as shown in Figure 2A and then exhibited a downhill trend for the rest of the 3 phases, reaching a low of 0.63% in 2016-2018. As shown in Figure 2B, for the P_2 of RA, there is only one significant increase of more than 0 in all 4 phases: 0.06 in 1951-1955 (Phase 1). This observation indicates that the popularity of RA in 1951-1955 increased by 6% compared to that in 1945-1950. It can also be observed from Figure 2C that the P_3 of RA was more than 5% during 1951-1980 and reached its maximum in Phase 1 (25%, 1960-1965), indicating that one quarter of the papers studying RA were published in the top 20 journals in the aspirin domain in Phase 1. In the next 3 phases, the P_3 peaked twice, in Phase 2 (1971-1975) and Phase 3 (2001-2005), possibly relating to the discovery of the mechanism of anti-inflammatory and RA-induced cardiovascular diseases. Similar to P_1, as shown in Figure 2D, the CI of RA peaked in 1956-1960 (40.44%), then declined to 1.02% in 2016-2018, indicating that around 40% of authors in Phase 1 were studying RA, but only about 1.02% authors still worked on the same disease in Phase 4.

In summary, in Phase 1, the P_1, P_2, P_3, and CI of RA reached their maxima, or showed a significant increase, indicating that RA was the disease upon which most research was focused in the aspirin domain at this time. However, the value of these indicators showed profound declines in the next 3 phases, which means that aspirin was studied in relation to other diseases and is thus an ideal example of drug repurposing.

Figure 2. The 4 entitymetric indexes of the biomedical entity “Rheumatoid Arthritis” over time. The background colors indicate the 4 phases of aspirin research.



Scientific Basis for Repurposing

As shown in Figure 3, there are 9 top biomedical entities in the aspirin domain that reached their maximum P_1 in Phase 2, including 3 diseases (“asthma”; “hypersensitivities, drug”; and “ulcer, gastric”) and 6 drugs (indomethacin, acetaminophen, dipyridamole, vitamin F, adenosine, and prostacyclin). The 3 diseases can all be side effects of aspirin, while the 6 drugs can be divided into 3 categories: (1) competitors of aspirin, that is, indomethacin and acetaminophen, which are analgesic and antipyretic drugs, respectively, with fewer side effects; (2) the antiplatelet drug dipyridamole; and (3) precursor substances in the pathway of the mechanism of action of aspirin (vitamin F, adenosine, and prostacyclin). In contrast with RA, the P_1 of these biomedical entities increased from Phase 1, peaked in Phase 2, and then decreased, indicating that the side effects and mechanisms of aspirin were studied in detail in Phase 2. The P_1 of indomethacin in 1976-1980 (16.75%) was the highest among these 9 entities in Phase 2, and vitamin F in 1981-1985 (11.19%) ranked second.

Figure 3. The Popularity Index (P_1) of the biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the 4 phases of aspirin research.

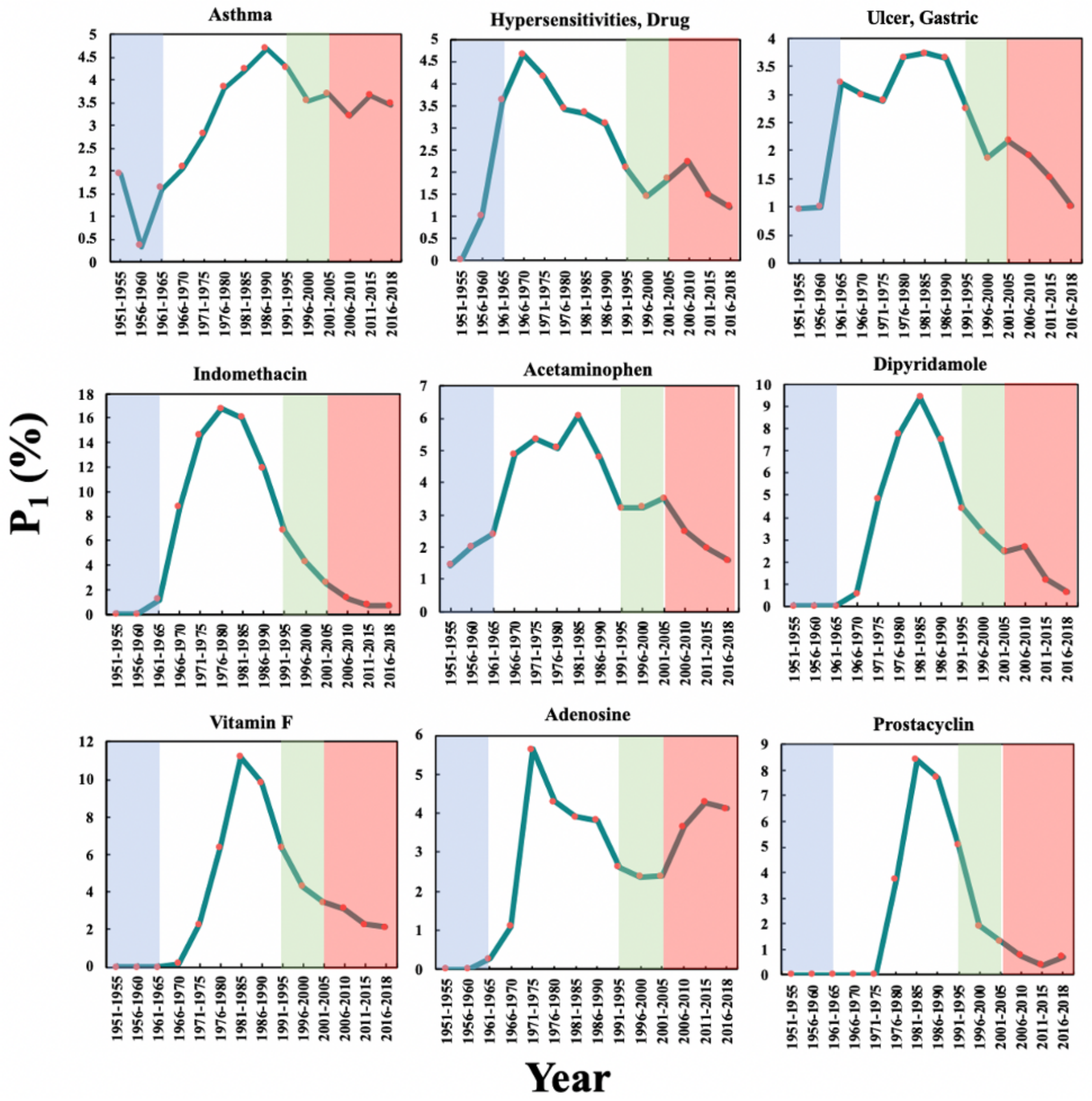


Figure 4 shows the P_2 of these 9 biomedical entities in the aspirin domain over time. The P_2 of the 3 side effects had a significant increase of more than zero in Phase 2, indicating that interest in the side effects of aspirin increased sharply: 1961-1965 and 1976-1980, for "asthma"; 1961-1965 for "hypersensitivities, drug"; and 1961-1965 for "ulcer, gastric." The time periods in which the P_2 of the 6 drugs showed significant increases are generally later than those for the side effects, such as 1971-1975 for indomethacin and 1981-1985 for prostacyclin. This observation indicates that the discovery and in-depth study of side effects may have positive effects on the discovery of the mechanism of action of aspirin as well as the development of alternatives with fewer side effects.

Figure 5 shows the P_3 of these 9 biomedical entities in the aspirin domain, demonstrating a feature common to all 9 entities: a gradual decline with a fluctuation in P_3 after reaching a maximum in Phase 1 or Phase 2. The highest initial P_3 values of "hypersensitivities, drug" and "ulcer, gastric" occurred in Phase 1, revealing that both side effects had been taken seriously by researchers in Phase 1. The P_3 of "hypersensitivities, drug" in 1956-1960 (33.33%) was higher than that of RA in 1956-1960 (25.00%). In 2011-2015, the P_3 of only 2 entities are over the 5% threshold: 5.82% for adenosine and 10.00% for prostacyclin. In the aspirin domain, papers studying these 2 entities published in the top 20 journals comprised more than 5% of papers published in all of the journals indexed by the PubMed in 2011-2015. This observation indicates that the 2 entities were still important foci of research in the aspirin domain.

It can be observed from Figure 3, Figure 5, and Table 3 that P_3, on average, achieved its maxima 10.7 years earlier than P_1. In particular, for “hypersensitivities, drug” and “ulcer, gastric,” the intervals can be as long as 20 years. This observation indicates that P_3 can reflect an early sign of academic interest into biomedical entities, a phenomenon that could be potentially valuable for tracking the research frontiers of a drug.

Figure 4. The Promising Index (P_2) of the biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the 4 phases of aspirin research.

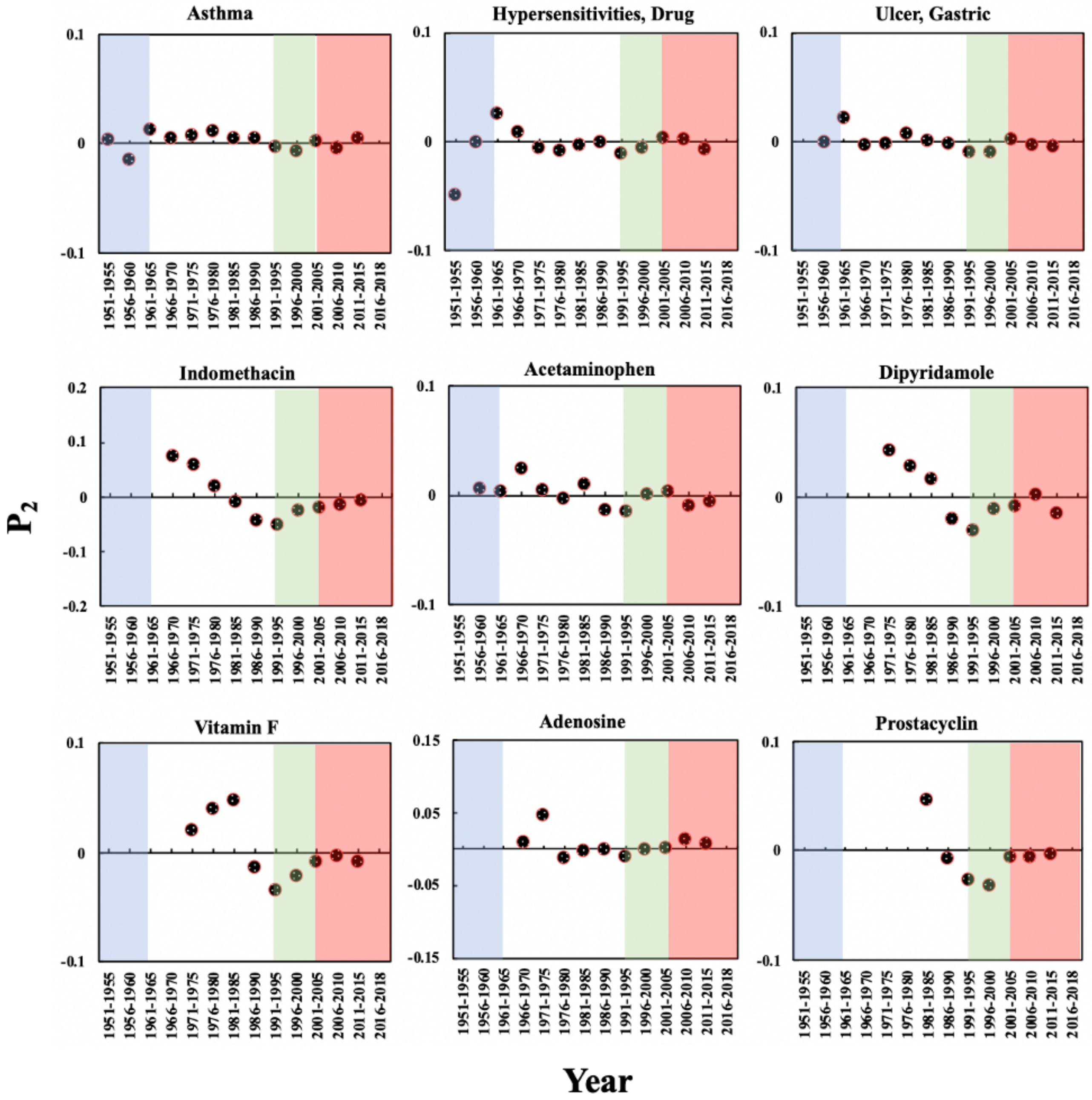


Figure 5. The Prestige Index (P_3) of the biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the 4 phases of aspirin research.

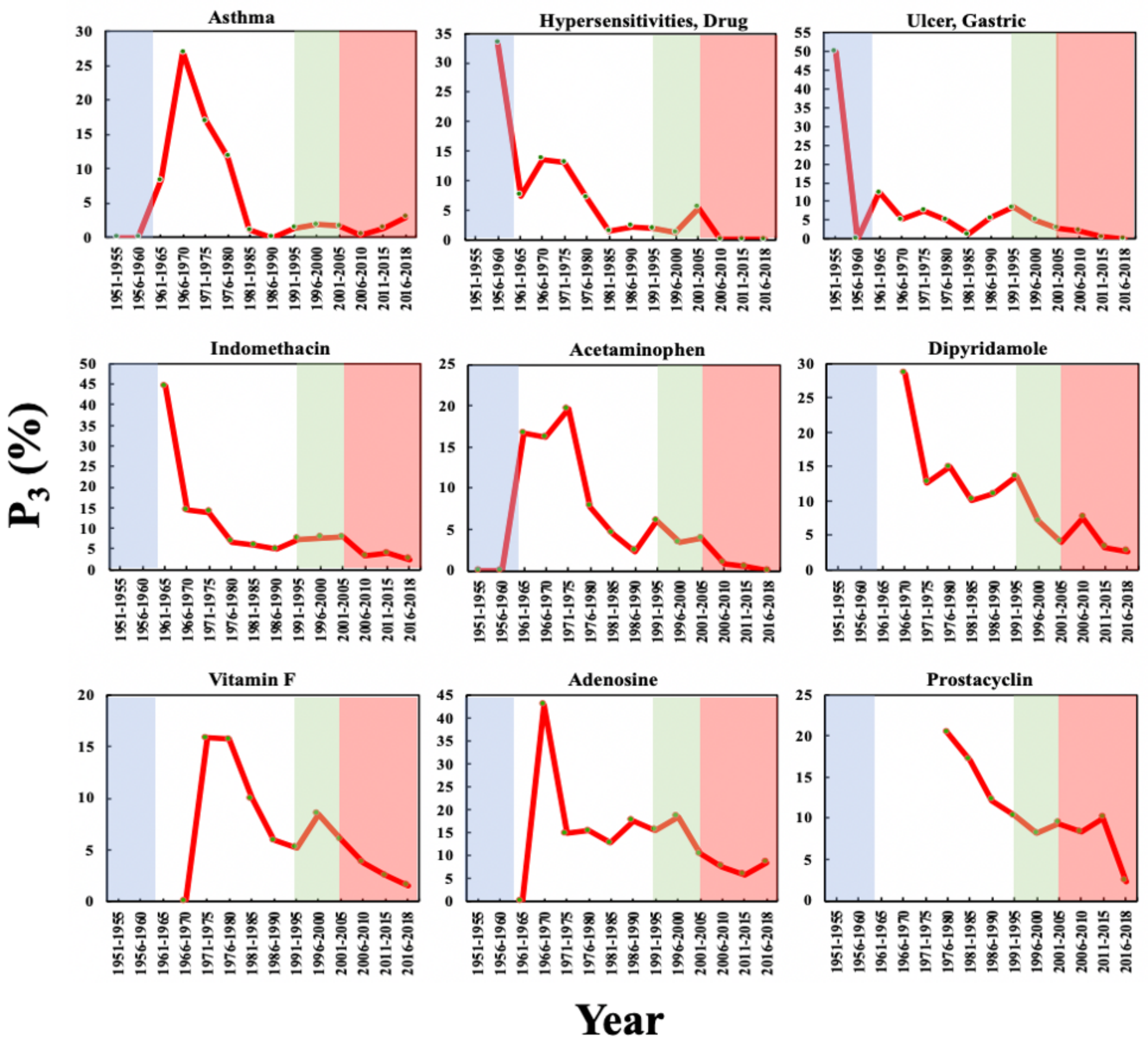


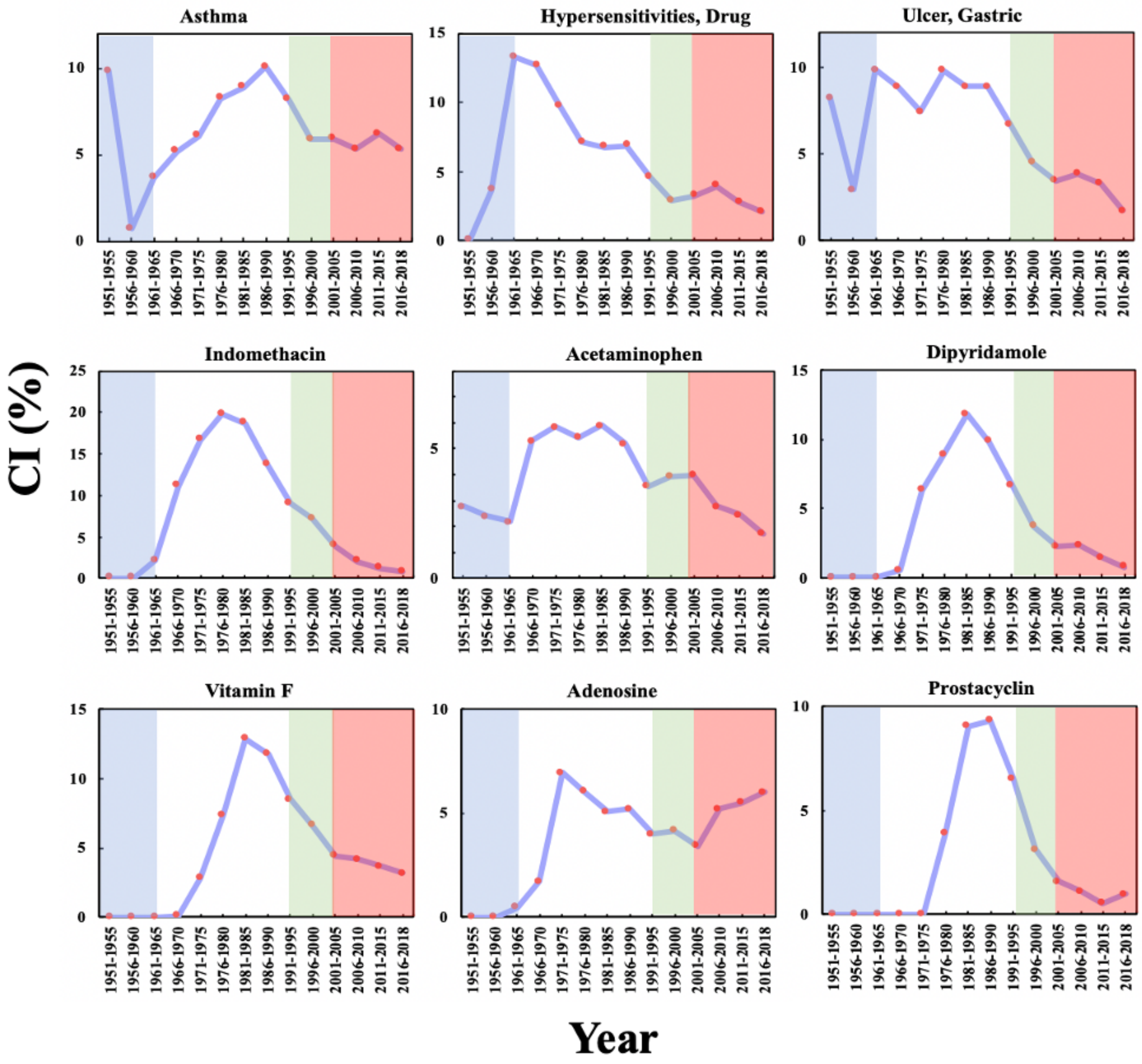
Table 3. Intervals between the time periods of the maxima of P_1 and P_3.

Biomedical entity	Time period of the maximum of P_1 (T1)	Time period of the maximum of P_3 (T2)	T1-T2 (years)
Asthma	1986-1990	1966-1970	20
Hypersensitivities, drug	1966-1970	1956-1960	10
Ulcer, gastric	1976-1980	1956-1960	20
Indomethacin	1976-1980	1961-1965	15
Acetaminophen	1981-1985	1971-1975	10
Dipyridamole	1981-1985	1966-1970	15
Vitamin F	1981-1985	1971-1975	10
Adenosine	1971-1975	1966-1970	5
Prostacyclin	1981-1985	1976-1980	5

The results of the CI of these 9 biomedical entities in the aspirin domain are presented in Figure 6, which shows that the CIs for these biomedical entities have similar trends to those of P_1 over time. Among all 9 biomedical entities during 1951-2018, indomethacin achieved the highest maximum CI in 1976-1980 (19.79%), indicating that it became a strong competitor to aspirin as an analgesic agent in Phase 2. This result also demonstrates that during the last 5-year period (2011-2015), the CIs of only 2 of the 9 entities were >5%, indicating that the 2 entities were still the subject of research by a considerable number of scientists (>2230) in the aspirin research community in 2011-2015. The 2 biomedical entities include “asthma” (6.21%) and adenosine (5.50%).

Based on the observation of P3C in Phase 2 and previous studies on aspirin [34,35], we can conclude that, on one hand, the in-depth investigation of the side effects and mechanism of action of aspirin provided the knowledge basis and research direction for drug repurposing. On the other hand, due to the market competition from other drugs, as well as the serious side effects, pharmaceutical companies attempted to discover new indicators for aspirin, in order to maintain the sales volume of aspirin.

Figure 6. Collaboration Index (CI) of the biomedical entities on the pharmacological mechanisms and side effects of aspirin over time. The background colors show the 4 phases of aspirin research.



Repurposing Aspirin for Cardiovascular-Related Diseases

In Phase 3, 5 top biomedical entities comprising 4 diseases and 1 drug reached their maximum P_1, as shown in Figure 7A. The 4 diseases were all cardiovascular-related, including “coronary disease” (P_1 of 18.88% in 1996-2000), “cerebral ischemia” (P_1 of 2.57% in 1996-2000), “intracranial vascular disorder” (P_1 of 5.73% in 1991-1995), and “ischemic heart disease” (P_1 of 3.01% in 1996-2000). Compared with Figures 2 and 3, the P_1 of the previous 10 biomedical entities that peaked in the Phase 1 or Phase 2 were considerably lower than that of coronary disease, indicating that cardiovascular-related disease was the focus of the aspirin domain in that time. Coronary disease is often referred as ischemic heart disease and is the most common cardiovascular-related disease worldwide; similarly, cerebral ischemia and intracranial vascular disorder represent the same condition, commonly known as stroke. These conditions were reportedly the first and second most common causes of death worldwide in the early 21st century [74]. The demand for the prevention and treatment of such fatal diseases could be one of the factors driving the repurposing of aspirin for cardiovascular-related diseases.

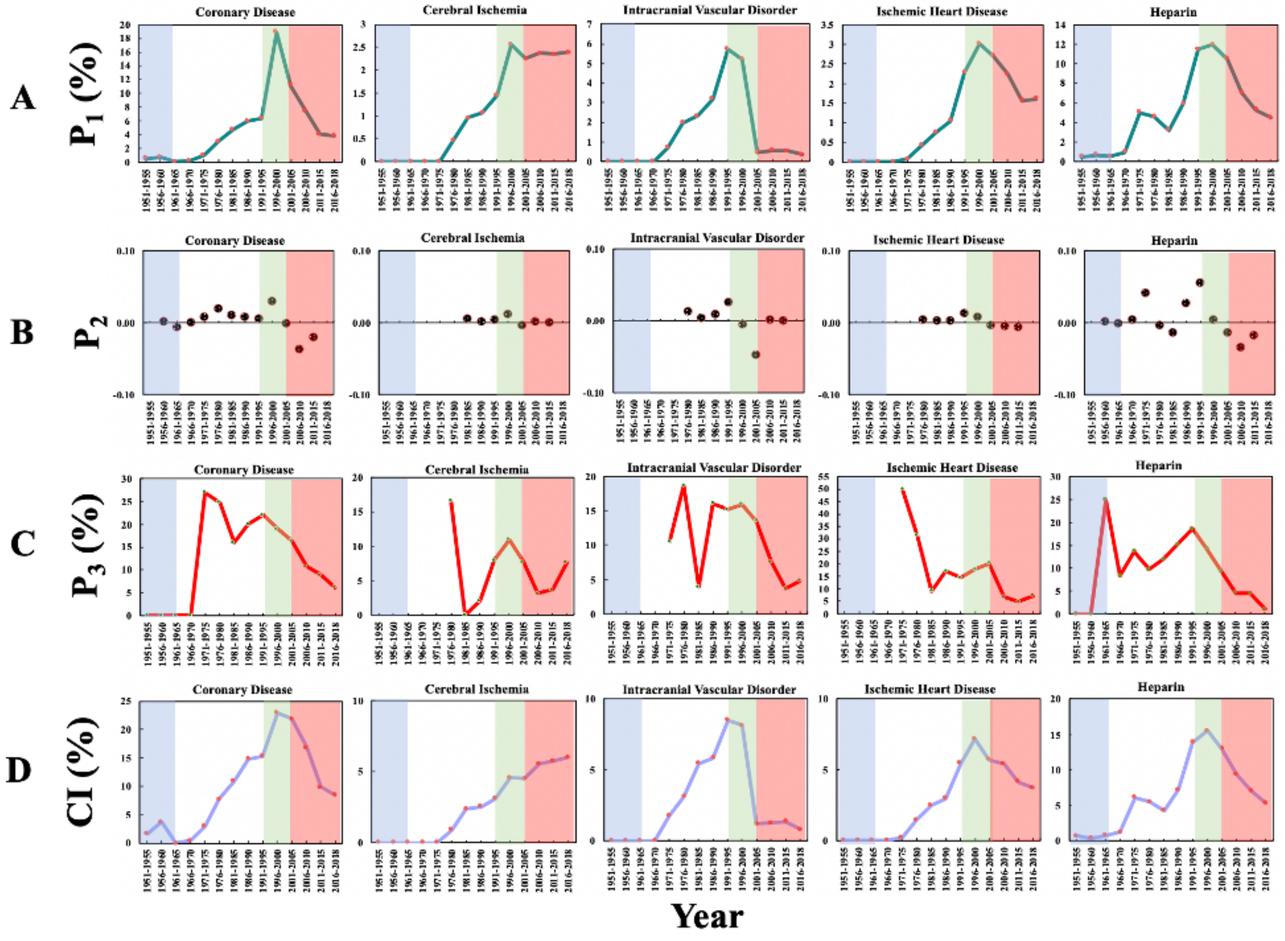
The only drug that reached its maximum P_1 in Phase 3 is heparin (11.92% in 1996-2000). As one of the most common anticoagulant drugs, heparin has always been the reference drug for repurposing aspirin to treat cardiovascular-related diseases, which could be the reason for the increase in the academic interest in heparin in the aspirin domain. There was another peak of heparin in Phase 2 (5.03%, 1971-1975), which could be related to an increase in research into the mechanisms of the antiplatelet effect of aspirin in Phase 2.

Figure 7B shows the changes in P_2 of these 5 biomedical entities over time. All 5 biomedical entities demonstrated a significant increase in Phase 3. “Coronary disease” and “cerebral ischemia” increased in 1991-1995, and “intracranial vascular disorder”, “ischemic heart disease,” and heparin increased in 1991-1995. The P_2 of the 2 entities also showed significant increases in Phase 2, consistent with the fact that aspirin was clinically used for coronary disease before the discovery of its antiplatelet effect: 0.02 in 1976-1980 for “coronary disease” and 0.10 in 1971-1975 for heparin.

The pattern of P_3 for these 5 entities over time is displayed in Figure 7C. All 5 biomedical entities reached their maxima in Phase 2, earlier than the maximum of P_1. “Coronary disease” reached a maximum in 1971-1975, and heparin reached a maximum in 1961-1965. The difference from the previous phases is that the P_3 of these 5 biomedical entities peaked again in Phase 3. For instance, “coronary disease” peaked in 1991-1995, and heparin peaked in 1991-1995, indicating that these biomedical entities were important topics of research in both Phase 1 and Phase 3.

Figure 7D shows the CI of the 5 biomedical entities during 1951-2018, in which the CI demonstrated a dynamic trajectory very similar to that of P_1. The maximum of “coronary disease” in Phase 3 is highest at 22.91% in 1996-2000, indicating that “coronary disease” attracted the greatest share of the authors in the aspirin domain. “Coronary disease” and “cerebral ischemia” in Phase 4 and heparin in Phases 2 and 4 surpassed the threshold value of 5%. The CI of “cerebral ischemia” steadily grew after Phase 3, showing a different trend from the other 4 biomedical entities, which increased in Phase 1 and Phase 2, peaked in Phase 3, and then dramatically decreased. This observation may illustrate that “cerebral ischemia,” unlike the other biomedical entities, is still increasing in popularity and collaboration, so additional increases are still expected.

Figure 7. The 4 entitymetric indexes of the biomedical entities on cardiovascular diseases in the aspirin domain over time. The background colors show the 4 phases of aspirin research.



Repurposing Aspirin for Other Diseases

In Figure 8, there are 15 biomedical entities that reached their maximum P_1 in Phase 4. Unlike the previous phases, most of the biomedical entities were genes and can be divided into 3 categories according to the diseases to which they are related: (1) inflammatory-related genes (eg, COX-2, LPLA2, and TNF-α), (2) cardiovascular-related genes (eg, COX-1, CD143, plasminogen, LDLCQ3, GPIIb, $P2Y_{12}$, and tPA), and (3) cancer-related genes (eg, TNFa, COX-2, COX-1, and LPLA2). These observations indicate that aspirin was actively studied for these 3 aspects of diseases from the perspective of genes in Phase 4. In particular, the maximum P_1 of COX-2 was the highest among these 15 biomedical entities at 21.97% in 2001-2005, revealing that COX-2 was considered to be very important in the aspirin domain at that time.

Figure 8 also shows that the P_1 of 2 diseases peaked in Phase 4. One is “diabetes,” whose P_1 in 2006-2010 was 6.83%. In fact, as early as 1875, Ebstein and Müller [75] discovered that aspirin had the effect of lowering blood glucose levels. Inspired by this observation, scientists have since been trying to use aspirin for the treatment of diabetes [75]. There are several peaks in the P_1 of “diabetes” in previous phases. In the 21st century, it has been recommended that patients with diabetes who have an increased risk of cardiovascular disease take aspirin as a primary preventative [5,76]; this could be the reason why the academic interest in “diabetes” in the aspirin domain increased again. The other disease is “carcinomas, colorectal.” Its P_1 peaked in 2001-2005 and then increased significantly after a small decline in 2006-2010, a pattern which is very different from other diseases in the aspirin domain. Repurposing aspirin for the treatment of colorectal carcinomas appears to be a focus of research in the aspirin domain today. The P_1 of 3 drugs also peaked in Phase 4, including the antiplatelet drugs clopidogrel and ticlopidine, which are competitors of aspirin as antiplatelet drugs [35], and warfarin, which is an anticoagulation drug that is similar to heparin and has been found to be superior to aspirin for secondary prevention of ischemic stroke with nonvalvular atrial fibrillation [77,78].

Figure 8. The Popularity Index (P_1) of the biomedical entities on repurposing aspirin for other diseases over time. The background colors show the 4 phases of aspirin research.

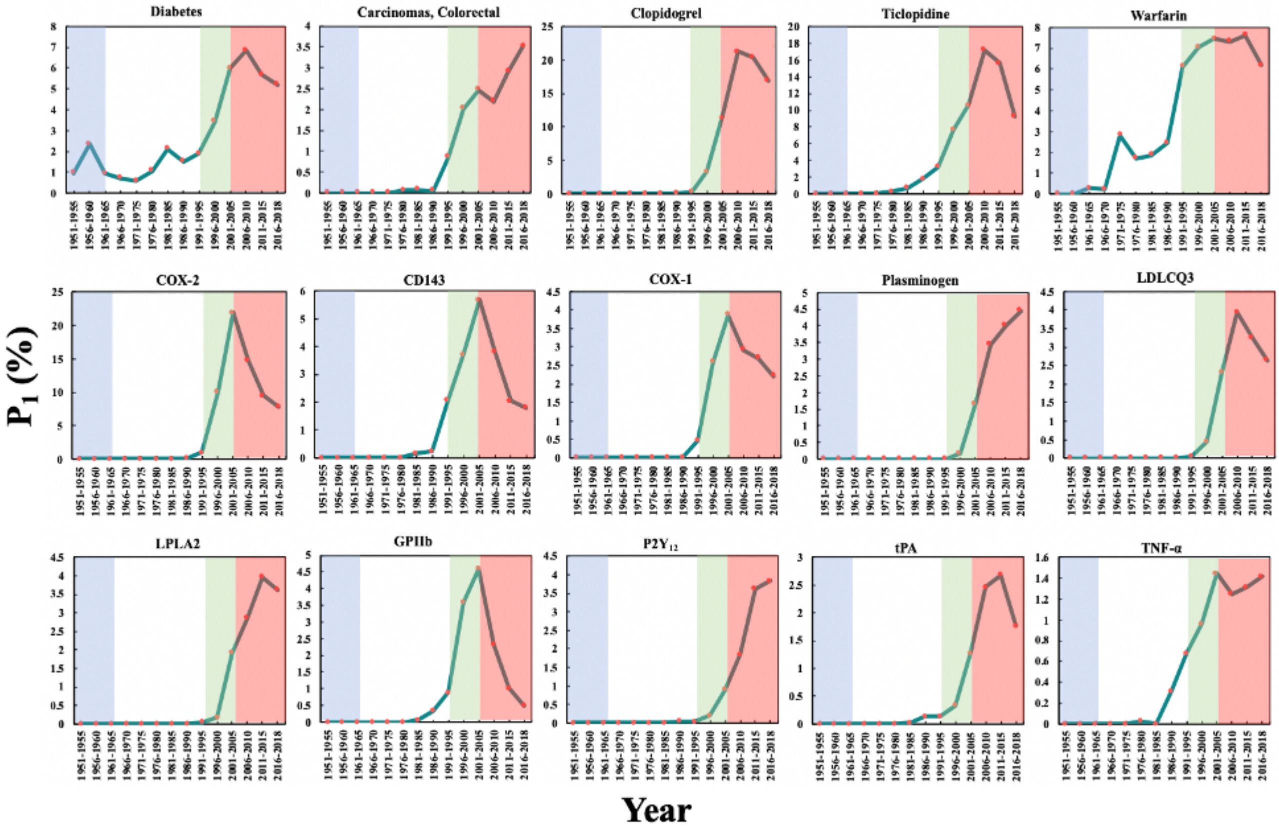


Figure 9 presents the changes in P_2 of these 15 biomedical entities over time. All of the genes demonstrate an increase of more than 0 in Phase 4. Unlike these genes, the diseases and drugs showed several significant increases of more than 0 in different phases, which reflects a longer history of research in the aspirin domain. For example, the increases occurred in 1956-1960, 1996-2000, and 2001-2005 for “diabetes”; 1996-2000, 2001-2005, and 2006-2010 for clopidogrel; and 1971-1975 and 1991-1995 for warfarin.

The changes in P_3 of these 15 biomedical entities over time are shown in Figure 10, from which we can make two observations. First, the P_3 of these biomedical entities demonstrated that the time period of the maximum of P_3 was much earlier than that of the maximum of P_1. Second, unlike the biomedical entities noted in previous sections, the diseases and drugs had ≥2 significant peaks in different phases. For instance, “diabetes” had peaks of 42.86% in 1956-1960, 25.00% in 1971-1975, and 14.22% in 1996-2000, and “carcinomas, colorectal” had peaks of 33.33% in 1981-1985, 15.91% in 1991-1995, and 14.15% in 2006-2010. These numbers indicate that these entities attracted considerable interest in the field of aspirin research and high-impact papers on these conditions were published. However, the genes usually had only one peak in P_3 in Phase 3 or 4, illustrating that these genes are relatively new topics in the aspirin domain.

The CI data for these 15 biomedical entities are presented in Figure 11, which shows that the maximum CI for COX-2 is the highest, at 34.37%, in 2001-2015, denoting that COX-2 was the focus of aspirin research in Phase 4; the research and development of Vioxx, a selective COX-2 inhibitor with fewer side effects, may be one of the reasons [79]. The CI of 2 drugs, clopidogrel (25.54% in 2001-2015) and ticlopidine (20.74% in 2006-2010), reveals fierce competition between aspirin and these alternative antiplatelet drugs. This competition could have driven the repurposing of aspirin for other diseases, especially cancers, that have an urgent demand for effective treatment.

Figure 9. The Promising Index (P_2) of the biomedical entities on repurposing aspirin for other diseases over time. The background colors show the 4 phases of aspirin research.

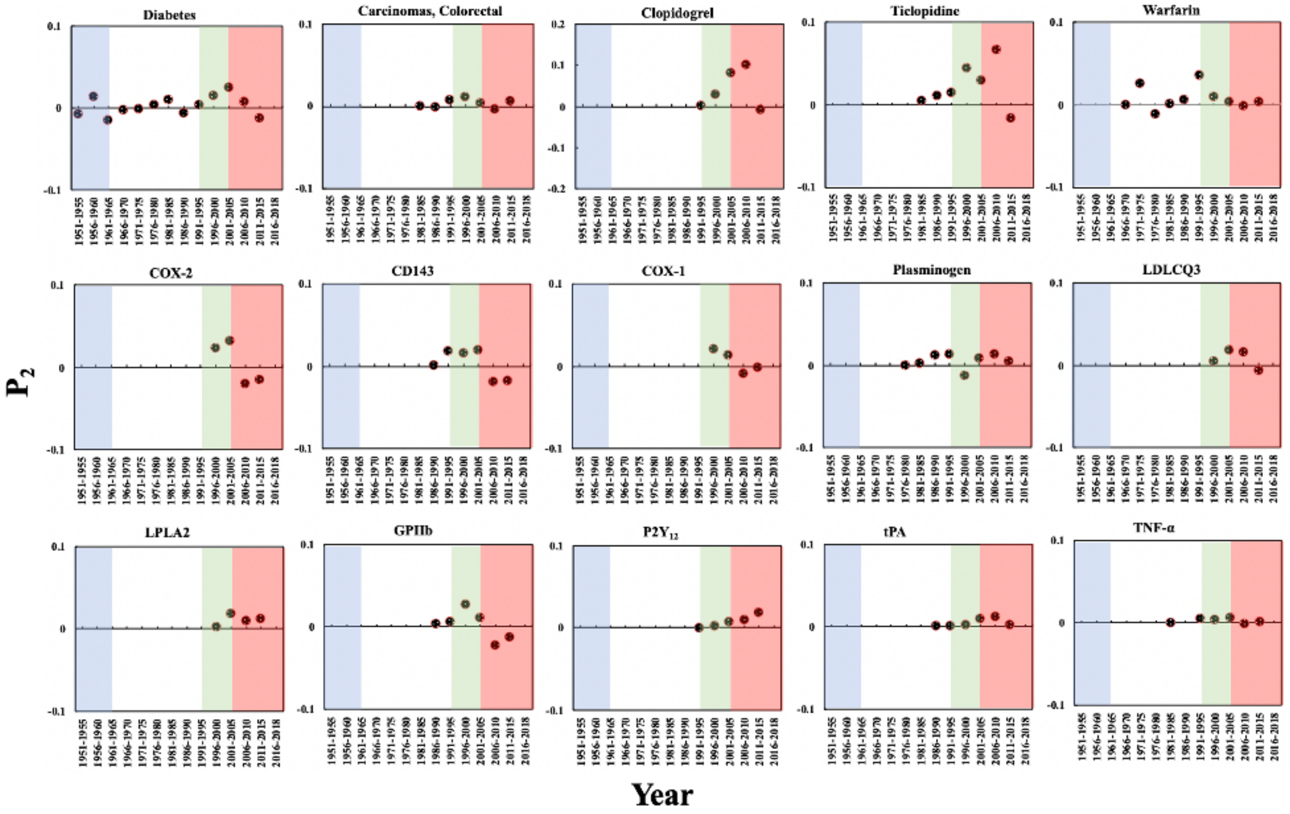


Figure 10. The Prestige Index (P_3) of the biomedical entities on repurposing aspirin for other diseases over time. The background colors show the 4 phases of aspirin research.

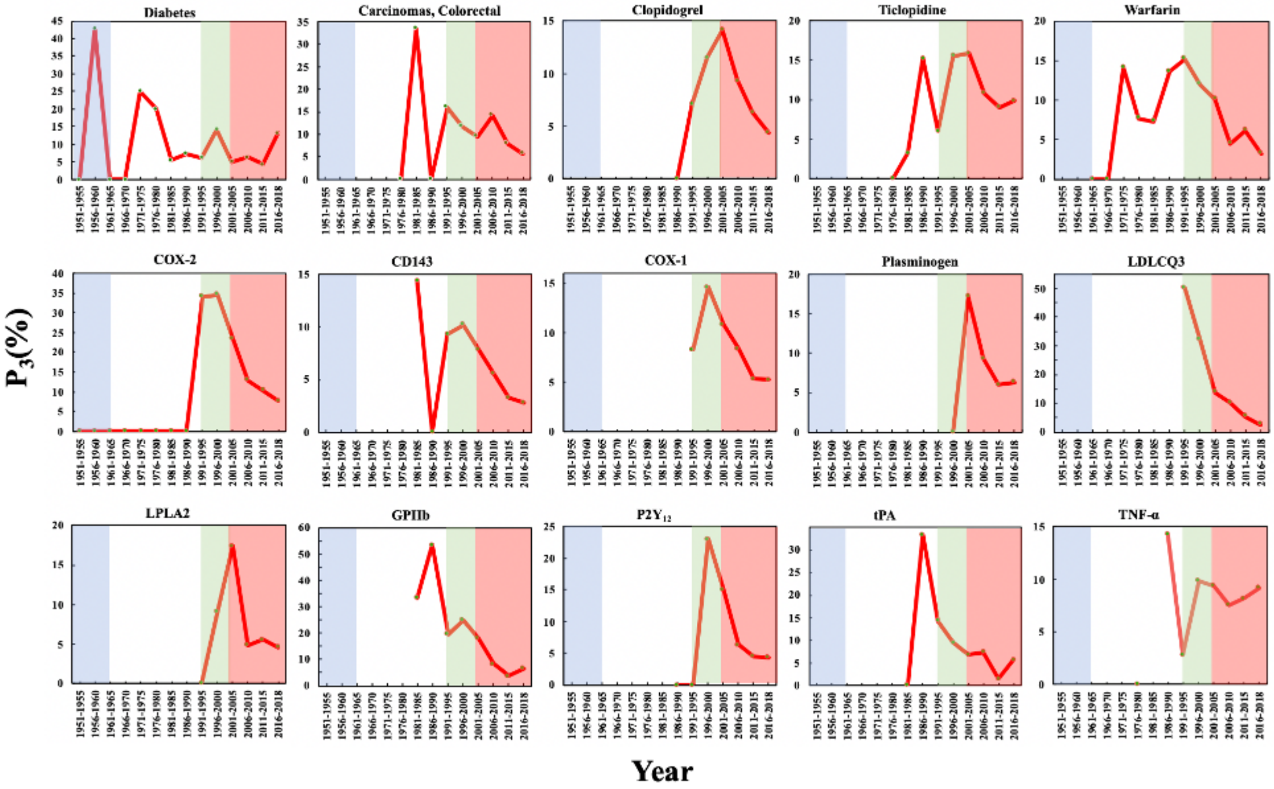
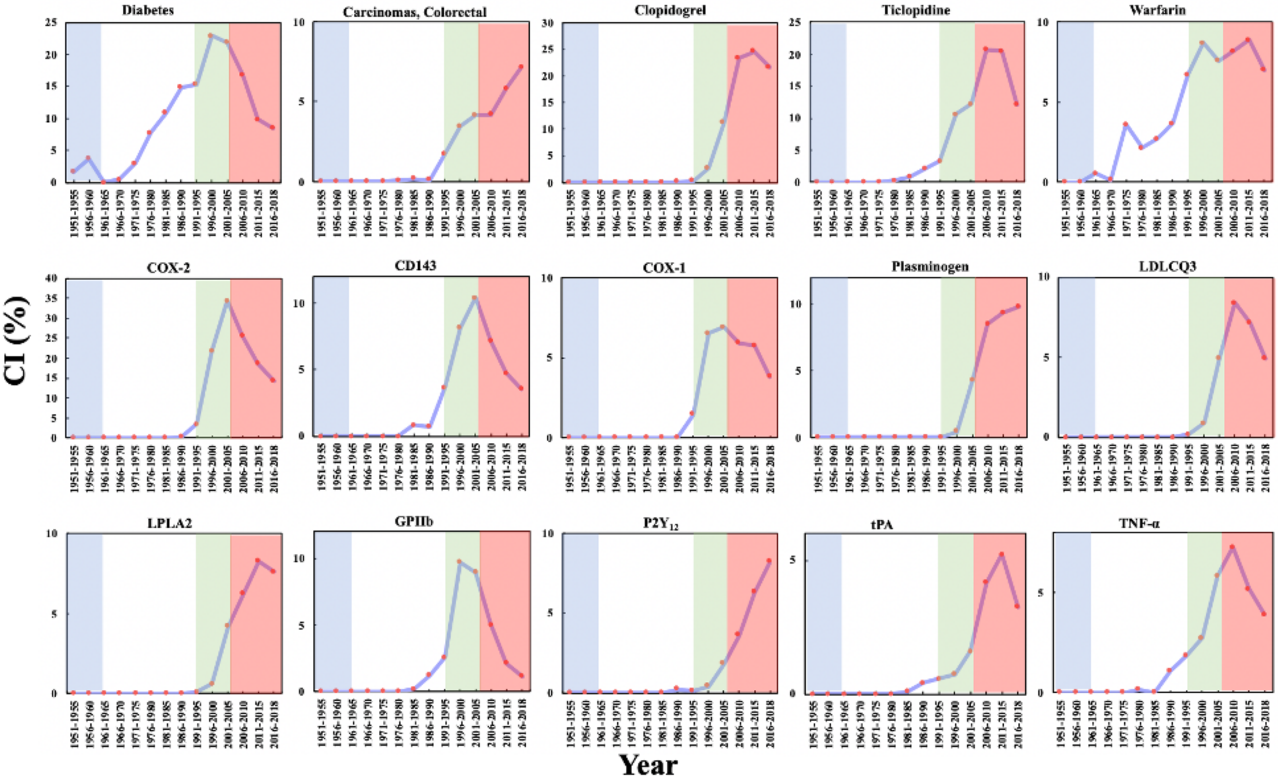


Figure 11. The Collaboration Index (CI) of the biomedical entities on repurposing aspirin for other diseases over time. The background colors show the 4 phases of aspirin research.



Discussion

Principal Findings

This study examines drug repurposing from the perspective of the evolution of biomedical entities, using aspirin as the study subject. It is of paramount importance for drug discovery to identify the factors that drive repurposing as well as potential patterns among biomedical entities in various phases of the drug research timeline. The main contribution of this paper is twofold. First, we proposed 4 entitymetric indices (P3C) to quantify changes in academic interest in biomedical entities and to reveal the granular process of drug repurposing. Second, we divided aspirin research into 4 phases, including original use (1951-1960), in-depth studies of pharmacological mechanisms and side effects (1961-1990), repurposing for cardiovascular-related diseases (1991-2000), and repurposing for other diseases (2001-2018), taking into consideration 3 granular perspectives—disease, drug, and gene—that contribute to a comprehensive understanding of the features of the repurposing process.

Our entitymetric results indicate that aspirin is representative of the process of drug repurposing. The research findings can be summarized as follows. In Phase 1, aspirin was routinely used to ease pain, fever, and inflammation and was often used in the treatment of RA [34], with a P3C that peaked in 1951-1960. Despite the widespread use of aspirin, at this stage, its mechanism of action was not well understood [34]. In Phase 2, the side effects and mechanisms of actions of aspirin were studied extensively, as shown by the maxima of P_1 and CI, as well as a significant increase in P_2 for the relevant biomedical entities in 1961-1990. The anti-platelet effect [70], inhibition of prostaglandin synthesis [71], and acetylation effect on the enzyme cyclo-oxygenase [72] were uncovered. These discoveries provided a solid knowledge foundation for the successful repurposing of aspirin. The highest P_1 in 1961-1990 was for indomethacin (16.75%), denoting fierce competition with aspirin for its original use. This could be one of the factors contributing to the repurposing of aspirin.

In Phase 3, aspirin was successfully used for several cardiovascular-related diseases because of its antiplatelet effects [80]. The related diseases and drugs achieved their highest values of P_1 and CI as well as significant increases in P_2 in 1991-2000. As these diseases are the most common diseases worldwide, according to data from the World Health Organization [74], the demand for the prevention and treatment of fatal diseases is also another potential factor driving drug repurposing. In the last phase, there was a large number of studies suggesting the use of aspirin for other diseases, especially colorectal cancer [36]. The greatest difference from previous phases is that aspirin was studied at the genetic level. Ten genes reached their maxima of P_1 and CI as well as an apparent increase in P_2 in 2001-2018. This observation could indicate that the development of modern science and technology, such as gene sequencing, molecular simulation, and deep learning, accelerates the process of drug repurposing of aspirin. Meanwhile, 2 fatal diseases — diabetes and colorectal carcinoma — as well as 3 competitive drugs of aspirin as an antiplatelet agent — clopidogrel, ticlopidine, and warfarin (an anticoagulant and competitor with aspirin for stroke prevention) — also had peak P_1 and CI values and a great increase in P_2.

Methodologically, in this study, we developed 4 entitymetrics and demonstrated how to use them to investigate the process of drug repurposing. The results demonstrate that the maxima of P_1, P_3, and CI are closely associated with the different phases of research of aspirin repurposing. The P_1 and CI metrics can indicate dynamic trends in academic interest in a given biomedical entity over a long time period. For instance, long-lasting increases in P_1 and CI signal interest in repurposing, while P_2 is more sensitive to immediate changes in academic interest in a specific biomedical entity, since it takes into consideration data from the two most recent periods. Moreover, P_3 can reflect a research focus far earlier than the other 3 indices, which means that a continuously high P_3 may be valuable as an early signal of the emergence and transfer of research topics in drug research. If P_3 does indeed have predictive power, it could be due to the involvement of top domain experts in the peer review of manuscripts in top journals with high impact factors [81,82]. Additionally, due to their easy implementation and interpretability, these indices can be applied in multiple domains, such as drug assessment, drug discovery, and pharmacovigilance.

Limitations and Future Directions

There are several limitations in the current paper. First, the data included in our analysis are limited to articles indexed in PubMed. Some real-world data, such as electronic health records, clinical trials, and social media, in which aspirin and its related biomedical entities were mentioned, should be included. In our future work, we will use different types of data sources for studying drug repurposing and take into account other entities related to drugs, including other biomedical entities, such as pathways, proteins, and cells, and non-biomedical entities, such as authors, institutions, and countries. The landscape of collaborations between academic institutions and pharmaceutical companies could affect the drug repurposing process. Second, there are several ways of measuring the impact of a journal, such as the impact factor and relative citation ratio. Third, this study mainly focused on investigating the repurposing journey of aspirin, but we did not test whether it can be used to predict future drug repurposing. In future studies, we will evaluate the different impact measures of a journal and choose a proper measure better fitted to the chosen drug. Furthermore, we will also aim to test the proposed metrics on other drugs to understand their repurposing journeys (eg, metformin) to see whether generalized patterns exist in different repurposing processes.

Acknowledgments

This study was supported by the Major Project of the National Natural Science Foundation of China (71673211). The support provided by the China Scholarship Council (CSC) during a visit by Xin Li to Indiana University Bloomington (No. 201806270047) is acknowledged. This work was also partly supported by the Bio-Synergy Research Project (NRF-2013M3A9C4078138) of the Ministry of Science, ICT, and Future Planning through the National Research Foundation. The authors are also grateful to the anonymous referees and editors for their invaluable and insightful comments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Top 20 journals related to aspirin research.
[DOCX File , 13 KB - medinform_v8i6e16739_app1.docx]

Multimedia Appendix 2
Changes in the number of journals with aspirin-related publications during 1951-2018. The top 5 journals and their frequencies are indicated using heat maps for every 5-year period.
[DOCX File , 587 KB - medinform_v8i6e16739_app2.docx]

Multimedia Appendix 3
Table S2 summarizes the peaks of P_1, P_3, and CI as well as the increase in P_2 for all the top 30 bioentities. The details can be found in the supplementary information section, including Phase 2 (1961-1990, the scientific basis for repurposing), Phase 3 (1991-2000, repurposing aspirin for cardiovascular-related diseases), and Phase 4 (2001-2018, repurposing aspirin for other diseases).
[DOCX File , 26 KB - medinform_v8i6e16739_app3.docx]

References

1. Schneider G. Automating drug discovery. Nat Rev Drug Discov 2018 Feb;17(2):97-113. [doi: 10.1038/nrd.2017.232] [Medline: 29242609]
2. Parrish MC, Tan YJ, Grimes KV, Mochly-Rosen D. Surviving in the Valley of Death: Opportunities and Challenges in Translating Academic Drug Discoveries. Annu Rev Pharmacol Toxicol 2019 Jan 06;59:405-421. [doi: 10.1146/annurev-pharmtox-010818-021625] [Medline: 30208282]
3. Deftereos SN, Andronis C, Friedla EJ, Persidis A, Persidis A. Drug repurposing and adverse event prediction using high-throughput literature analysis. WIREs Syst Biol Med 2011 Feb 16;3(3):323-334. [doi: 10.1002/wsbm.147]
4. Cha Y, Erez T, Reynolds IJ, Kumar D, Ross J, Koytiger G, et al. Drug repurposing from the perspective of pharmaceutical companies. British Journal of Pharmacology 2017 May 18;175(2):168-180. [doi: 10.1111/bph.13798]
5. Strittmatter S. Overcoming Drug Development Bottlenecks With Repurposing: Old drugs learn new tricks. Nat Med 2014 Jun;20(6):590-591 [FREE Full text] [doi: 10.1038/nm.3595] [Medline: 24901567]
6. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med 2017 Apr 07;23(4):405-408 [FREE Full text] [doi: 10.1038/nm.4306] [Medline: 28388612]
7. Sheard S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2019 Jan;18(1):41-58 [FREE Full text] [doi: 10.1038/nrd.2018.168] [Medline: 30310233]
8. Fenig DM, McCullough A. Sildenafil in the treatment of erectile dysfunction. Aging Health 2007 Jun;3(3):295-303.
9. Singhal S, Mehta J, Desikan R, Ayers D, Roberson P, Eddlemon P, et al. Antitumor Activity of Thalidomide in Refractory Multiple Myeloma. N Engl J Med 1999 Nov 18;341(21):1565-1571. [doi: 10.1056/nejm199911183412102]
10. Lee A, Morley J. Metformin decreases food consumption and induces weight loss in subjects with obesity with type II non-insulin-dependent diabetes. Obes Res 1998 Jan;6(1):47-53 [FREE Full text] [doi: 10.1002/j.1550-8528.1998.tb00314.x] [Medline: 9526970]
11. Berstein LM. Metformin in obesity, cancer and aging: addressing controversies. Aging (Albany NY) 2012 May;4(5):320-329 [FREE Full text] [doi: 10.18632/aging.100455] [Medline: 22589237]
12. Pascual J, Rivas MT, Leira R. Testing the combination beta-blocker plus topiramate in refractory migraine. Acta Neurol Scand 2007 Feb;115(2):81-83. [doi: 10.1111/j.1600-0404.2006.00772.x] [Medline: 17212609]
13. Diener H, Tfelt-Hansen P, Dahlöf C, Láinez MJA, Sandrini G, Wang S, MIGR-003 Study Group. Topiramate in migraine prophylaxis--results from a placebo-controlled trial with propranolol as an active control. J Neurol 2004 Aug;251(8):943-950. [doi: 10.1007/s00415-004-0464-6] [Medline: 15316798]

14. Simsek M, Meijer B, van Bodegraven AA, de Boer NK, Mulder CJ. Finding hidden treasures in old drugs: the challenges and importance of licensing generics. Drug Discov Today 2018 Jan;23(1):17-21 [FREE Full text] [doi: 10.1016/j.drudis.2017.08.008] [Medline: 28867540]
15. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nat Commun 2017 Jul 12;8(1):16022 [FREE Full text] [doi: 10.1038/ncomms16022] [Medline: 28699633]
16. Xu M, Lee EM, Wen Z, Cheng Y, Huang W, Qian X, et al. Identification of small-molecule inhibitors of Zika virus infection and induced neural cell death via a drug repurposing screen. Nat Med 2016 Oct 29;22(10):1101-1107 [FREE Full text] [doi: 10.1038/nm.4184] [Medline: 27571349]
17. Mons B, van Haagen H, Chichester C, Hoen P, den Dunnen JT, van Ommen G, et al. The value of data. Nat Genet 2011 Mar 29;43(4):281-283. [doi: 10.1038/ng0411-281] [Medline: 21445068]
18. Gao Z, Fu G, Ouyang C, Tsutsui S, Liu X, Yang J, et al. edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. BMC Bioinformatics 2019 Jun 10;20(1):306 [FREE Full text] [doi: 10.1186/s12859-019-2914-2] [Medline: 31238875]
19. Chen B, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, et al. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 2010 May 17;11:255 [FREE Full text] [doi: 10.1186/1471-2105-11-255] [Medline: 20478034]
20. Hamilton W, Bajaj P, Zitnik M, Jurafsky D, Leskovec J. Embedding Logical Queries on Knowledge Graphs. In: the proceedings of 32nd International Conference on Neural Information Processing Systems (NIPS'18). 2018 Presented at: NIPS18; Dec 3 – Dec 8; Montreal, Canada p. 2030-2041.
21. Chang Y, Park H, Yang H, Lee S, Lee K, Kim TS, et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci Rep 2018 Jun 11;8(1):8857 [FREE Full text] [doi: 10.1038/s41598-018-27214-6] [Medline: 29891981]
22. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 2018 Sep 01;34(17):i821-i829 [FREE Full text] [doi: 10.1093/bioinformatics/bty593] [Medline: 30423097]
23. Jinha AE. Article 50 million: an estimate of the number of scholarly articles in existence. Learn. Pub 2010 Jul 01;23(3):258-263. [doi: 10.1087/20100308]
24. Ding Y, Stirling K. Data-driven Discovery: A New Era of Exploiting the Literature and Data. JDIS 2016 Nov 03;1(4):1-9. [doi: 10.20309/jdis.201622]
25. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30(1):7-18. [doi: 10.1353/pbm.1986.0087] [Medline: 3797213]
26. Swanson DR. Undiscovered Public Knowledge. The Library Quarterly 1986 Apr;56(2):103-118. [doi: 10.1086/601720]
27. Swanson DR. Two medical literatures that are logically but not bibliographically connected. J. Am. Soc. Inf. Sci 1987 Jul;38(4):228-233. [doi: 10.1002/(sici)1097-4571]
28. Cory KA. Discovering Hidden Analogies in an Online Humanities Database. Comput Hum 1997;31(1):1-12. [doi: 10.1023/A:1000422220677]
29. Ding Y, Song M, Han J, Yu Q, Yan E, Lin L, et al. Entitymetrics: measuring the impact of entities. PLoS One 2013;8(8):e71416 [FREE Full text] [doi: 10.1371/journal.pone.0071416] [Medline: 24009660]
30. Wang H, Ding Y, Tang J, Dong X, He B, Qiu J, et al. Finding complex biological relationships in recent PubMed articles using Bio-LDA. PLoS One 2011 Mar 23;6(3):e17243 [FREE Full text] [doi: 10.1371/journal.pone.0017243] [Medline: 21448266]
31. Song M, Han N, Kim Y, Ding Y, Chambers T. Discovering implicit entity relation with the gene-citation-gene network. PLoS One 2013;8(12):e84639 [FREE Full text] [doi: 10.1371/journal.pone.0084639] [Medline: 24358368]
32. Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C. Content-based citation analysis: The next generation of citation analysis. J Assn Inf Sci Tec 2014 Jun 06;65(9):1820-1833. [doi: 10.1002/asi.23256]
33. Kissin I, Edwin LB. Top Journals Selectivity Index: is it acceptable for drugs beyond the field of analgesia? Scientometrics 2011 May 5;88(2):589-597. [doi: 10.1007/s11192-011-0403-0]
34. Montinari MR, Minelli S, De Caterina R. The first 3500 years of aspirin history from its roots - A concise summary. Vascul Pharmacol 2019 Feb;113:1-8. [doi: 10.1016/j.vph.2018.10.008] [Medline: 30391545]
35. Bordons M, Bravo C, Barrigón S. Time-tracking of the research profile of a drug using bibliometric tools. J. Am. Soc. Inf. Sci 2004 Jan 16;55(5):445-461. [doi: 10.1002/asi.10397]
36. Gilligan MM, Gartung A, Sulciner ML, Norris PC, Sukhatme VP, Bielenberg DR, et al. Aspirin-triggered proresolving mediators stimulate resolution in cancer. Proc Natl Acad Sci U S A 2019 Mar 26;116(13):6292-6297 [FREE Full text] [doi: 10.1073/pnas.1804000116] [Medline: 30862734]
37. Volz J, Mammadova-Bach E, Gil-Pulido J, Nandigama R, Remer K, Sorokin L, et al. Inhibition of platelet GPVI induces intratumor hemorrhage and increases efficacy of chemotherapy in mice. Blood 2019 Jun 20;133(25):2696-2706. [doi: 10.1182/blood.2018877043] [Medline: 30952674]
38. Jackson S, Zhu B, Pfeiffer R, Liu Z, Gadalla S, Koshiol J. Aspirin may extend biliary tract cancer survival: Results from population-based cohort. Clin Res 2019:2333. [doi: 10.1158/1538-7445.AM2019-2333]

39. Baker NC, Ekins S, Williams AJ, Tropsha A. A bibliometric review of drug repurposing. Drug Discov Today 2018 Mar;23(3):661-672 [FREE Full text] [doi: 10.1016/j.drudis.2018.01.018] [Medline: 29330123]
40. Wang Z, Lachmann A, Keenan AB, Ma'ayan A. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. Bioinformatics 2018 Jun 15;34(12):2150-2152 [FREE Full text] [doi: 10.1093/bioinformatics/bty060] [Medline: 29420694]
41. Liu T, Hsieh Y, Chou C, Yang P. Systematic polypharmacology and drug repurposing via an integrated L1000-based Connectivity Map database mining. R Soc Open Sci 2018 Nov;5(11):181321 [FREE Full text] [doi: 10.1098/rsos.181321] [Medline: 30564416]
42. Kidnapillai S, Bortolasci CC, Udawela M, Panizzutti B, Spolding B, Connor T, et al. The use of a gene expression signature and connectivity map to repurpose drugs for bipolar disorder. World J Biol Psychiatry 2018 Aug 03:1-9. [doi: 10.1080/15622975.2018.1492734] [Medline: 29956574]
43. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 Jun 07;447(7145):661-678 [FREE Full text] [doi: 10.1038/nature05911] [Medline: 17554300]
44. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, et al. Use of genome-wide association studies for drug repositioning. Nat Biotechnol 2012 Apr 10;30(4):317-320. [doi: 10.1038/nbt.2151] [Medline: 22491277]
45. Ferrero E, Agarwal P. Connecting genetics and gene expression data for target prioritisation and drug repositioning. BioData Min 2018;11:7 [FREE Full text] [doi: 10.1186/s13040-018-0171-y] [Medline: 29881461]
46. Dakshanamurthy S, Issa NT, Assefnia S, Seshasayee A, Peters OJ, Madhavan S, et al. Predicting new indications for approved drugs using a proteochemometric method. J Med Chem 2012 Aug 09;55(15):6832-6848 [FREE Full text] [doi: 10.1021/jm300576q] [Medline: 22780961]
47. Li YY, An J, Jones SJM. A computational approach to finding novel targets for existing drugs. PLoS Comput Biol 2011 Sep;7(9):e1002139 [FREE Full text] [doi: 10.1371/journal.pcbi.1002139] [Medline: 21909252]
48. Leinung MC, Grasso P. [D-Leu-4]-OB3, a synthetic peptide amide with leptin-like activity, augments the effects of orally delivered exenatide and pramlintide acetate on energy balance and glycemic control in insulin-resistant male C57BLK/6-m db/db mice. Regul Pept 2012 Nov 10;179(1-3):33-38. [doi: 10.1016/j.regpep.2012.08.006] [Medline: 22960403]
49. Yang L, Agarwal P. Systematic drug repositioning based on clinical side-effects. PLoS One 2011;6(12):e28025 [FREE Full text] [doi: 10.1371/journal.pone.0028025] [Medline: 22205936]
50. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. PLoS One 2014;9(2):e87864 [FREE Full text] [doi: 10.1371/journal.pone.0087864] [Medline: 24505324]
51. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches. Bioinformatics 2018 Apr 01;34(7):1164-1173 [FREE Full text] [doi: 10.1093/bioinformatics/btx731] [Medline: 29186331]
52. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Brief Bioinform 2014 Sep;15(5):734-747. [doi: 10.1093/bib/bbt056] [Medline: 23933754]
53. Liang X, Li D, Song M, Madden A, Ding Y, Bu Y. Predicting biomedical relationships using the knowledge and graph embedding cascade model. PLoS One 2019;14(6):e0218264 [FREE Full text] [doi: 10.1371/journal.pone.0218264] [Medline: 31194807]
54. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. PLoS Comput Biol 2012;8(7):e1002574 [FREE Full text] [doi: 10.1371/journal.pcbi.1002574] [Medline: 22859915]
55. Fu G, Ding Y, Seal A, Chen B, Sun Y, Bolton E. Predicting drug target interactions using meta-path-based semantic network analysis. BMC Bioinformatics 2016 Apr 12;17:160 [FREE Full text] [doi: 10.1186/s12859-016-1005-x] [Medline: 27071755]
56. Wang Q, Mao Z, Wang B, Guo L. Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Trans. Knowl. Data Eng 2017 Dec 1;29(12):2724-2743. [doi: 10.1109/tkde.2017.2754499]
57. Mukhopadhyay S, Palakal M, Maddu K. Multi-way association extraction and visualization from biological text documents using hyper-graphs: applications to genetic association studies for diseases. Artif Intell Med 2010 Jul;49(3):145-154. [doi: 10.1016/j.artmed.2010.03.002] [Medline: 20382004]
58. Cameron D, Bodenreider O, Yalamanchili H, Danh T, Vallabhaneni S, Thirunarayan K, et al. A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. J Biomed Inform 2013 Apr;46(2):238-251 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.004] [Medline: 23026233]
59. Song M, Heo GE, Ding Y. SemPathFinder: Semantic path analysis for discovering publicly unknown knowledge. Journal of Informetrics 2015 Oct;9(4):686-703. [doi: 10.1016/j.joi.2015.06.004]
60. Lv Y, Ding Y, Song M, Duan Z. Topology-driven trend analysis for drug discovery. Journal of Informetrics 2018 Aug;12(3):893-905. [doi: 10.1016/j.joi.2018.07.007]
61. 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]

62. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004 Jan 01;32(Database issue):D258-D261 [FREE Full text] [doi: 10.1093/nar/gkh036] [Medline: 14681407]
63. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016 Jan 04;44(D1):D1075-D1079 [FREE Full text] [doi: 10.1093/nar/gkv1075] [Medline: 26481350]
64. He B, Tang J, Ding Y, Wang H, Sun Y, Shin JH, et al. Mining relational paths in integrated biomedical data. PLoS One 2011;6(12):e27506 [FREE Full text] [doi: 10.1371/journal.pone.0027506] [Medline: 22162991]
65. Williams RS, Lotia S, Holloway AK, Pico AR. From scientific discovery to cures: bright stars within a galaxy. Cell 2015 Sep 24;163(1):21-23 [FREE Full text] [doi: 10.1016/j.cell.2015.09.007] [Medline: 26406364]
66. Zhu Y, Song M, Yan E. Identifying Liver Cancer and Its Relations with Diseases, Drugs, and Genes: A Literature-Based Approach. PLoS One 2016;11(5):e0156091 [FREE Full text] [doi: 10.1371/journal.pone.0156091] [Medline: 27195695]
67. Lerchenmueller MJ, Sorenson O. Author Disambiguation in PubMed: Evidence on the Precision and Recall of Author-ity among NIH-Funded Scientists. PLoS One 2016;11(7):e0158731 [FREE Full text] [doi: 10.1371/journal.pone.0158731] [Medline: 27367860]
68. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, et al. BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. PLoS One 2016;11(10):e0164680 [FREE Full text] [doi: 10.1371/journal.pone.0164680] [Medline: 27760149]
69. Kissin I. What Can Big Data on Academic Interest Reveal about a Drug? Reflections in Three Major US Databases. Trends Pharmacol Sci 2018 Mar;39(3):248-257. [doi: 10.1016/j.tips.2017.12.005] [Medline: 29358009]
70. O'Brien J. EFFECTS OF SALICYLATES ON HUMAN PLATELETS. The Lancet 1968 Apr;291(7546):779-783. [doi: 10.1016/s0140-6736(68)92228-9]
71. Vane JR. Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs. Nat New Biol 1971 Jun 23;231(25):232-235. [doi: 10.1038/newbio231232a0] [Medline: 5284360]
72. Roth GJ, Stanford N, Majerus PW. Acetylation of prostaglandin synthase by aspirin. Proc Natl Acad Sci U S A 1975 Aug;72(8):3073-3076 [FREE Full text] [doi: 10.1073/pnas.72.8.3073] [Medline: 810797]
73. Kune GA, Kune S, Watson LF. Colorectal cancer risk, chronic illnesses, operations, and medications: case control results from the Melbourne Colorectal Cancer Study. Cancer Res 1988 Aug 01;48(15):4399-4404 [FREE Full text] [Medline: 3390835]
74. World Health Organization. The top 10 causes of death Internet URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2019-07-16]
75. Ebstein W, Müller J. Brenzkatechin in dem Urin eines Kindes. Archiv f. pathol. Anat 1875 Feb;62(4):554-560. [doi: 10.1007/bf01928660]
76. Pignone M, Alberts MJ, Colwell JA, Cushman M, Inzucchi SE, Mukherjee D, American Diabetes Association, American Heart Association, American College of Cardiology Foundation. Aspirin for primary prevention of cardiovascular events in people with diabetes. J Am Coll Cardiol 2010 Jun 22;55(25):2878-2886 [FREE Full text] [doi: 10.1016/j.jacc.2010.04.003] [Medline: 20579547]
77. van Walraven C, Hart RG, Singer DE, Laupacis A, Connolly S, Petersen P, et al. Oral anticoagulants vs aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. JAMA 2002 Nov 20;288(19):2441-2448. [doi: 10.1001/jama.288.19.2441] [Medline: 12435257]
78. Mohr J, Thompson J, Lazar R, Levin B, Sacco R, Furie K, et al. A Comparison of Warfarin and Aspirin for the Prevention of Recurrent Ischemic Stroke. N Engl J Med 2001 Nov 15;345(20):1444-1451. [doi: 10.1056/nejmoa011258]
79. Couzin J. Drug safety. Withdrawal of Vioxx casts a shadow over COX-2 inhibitors. Science 2004 Oct 15;306(5695):384-385. [doi: 10.1126/science.306.5695.384] [Medline: 15486258]
80. Sanmuganathan PS, Ghahramani P, Jackson PR, Wallis EJ, Ramsay LE. Aspirin for primary prevention of coronary heart disease: safety and absolute benefit related to coronary risk derived from meta-analysis of randomised trials. Heart 2001 Mar;85(3):265-271 [FREE Full text] [doi: 10.1136/heart.85.3.265] [Medline: 11179262]
81. Kissin I. Can a bibliometric indicator predict the success of an analgesic? Scientometrics 2010 Nov 30;86(3):785-795. [doi: 10.1007/s11192-010-0320-7]
82. Koenig MED. Determinants of expert judgement of research performance. Scientometrics 1982 Sep;4(5):361-378. [doi: 10.1007/bf02135122]

Abbreviations

CI: Collaboration Index
DR: drug repurposing
GWAS: genome-wide association study
P_1: Popularity Index
P_2: Promising Index
P_3: Prestige Index

P3C: the 4 entitymetric indicators for biomedical entities
RA: rheumatoid arthritis

Edited by G Eysenbach; submitted 19.10.19; peer-reviewed by J Yang, W Griffin, Y Li; comments to author 06.12.19; revised version received 08.01.20; accepted 31.03.20; published 16.06.20.

Please cite as:
Li X, Rousseau JF, Ding Y, Song M, Lu W
Understanding Drug Repurposing From the Perspective of Biomedical Entities and Their Evolution: Bibliographic Research Using Aspirin
JMIR Med Inform 2020;8(6):e16739
URL: https://medinform.jmir.org/2020/6/e16739
doi:10.2196/16739
PMID:32543442

Original Paper

Summarizing Complex Graphical Models of Multiple Chronic Conditions Using the Second Eigenvalue of Graph Laplacian: Algorithm Development and Validation

Syed Hasib Akhter Faruqui[1], MSc; Adel Alaeddini[1], PhD; Mike C Chang[1], MSc; Sara Shirinkam[2], PhD; Carlos Jaramillo[3], MD, PhD; Peyman NajafiRad[4], PhD; Jing Wang[5], MPH, PhD; Mary Jo Pugh[6], PhD

[1]Department of Mechanical Engineering, The University of Texas at San Antonio, San Antonio, TX, United States
[2]Department of Mathematics and Statistics, University of the Incarnate Word, San Antonio, TX, United States
[3]South Texas Veterans Health Care System, San Antonio, TX, United States
[4]Department of Information Systems and Cyber Security, The University of Texas at San Antonio, San Antonio, TX, United States
[5]School of Nursing, UT Health San Antonio, San Antonio, TX, United States
[6]VA Salt Lake City Health Care System, Salt Lake City, UT, United States

Corresponding Author:
Adel Alaeddini, PhD
Department of Mechanical Engineering
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX, 78249
United States
Phone: 1 210 458 8747
Email: adel.alaeddini@utsa.edu

Abstract

Background: It is important but challenging to understand the interactions of multiple chronic conditions (MCC) and how they develop over time in patients and populations. Clinical data on MCC can now be represented using graphical models to study their interaction and identify the path toward the development of MCC. However, the current graphical models representing MCC are often complex and difficult to analyze. Therefore, it is necessary to develop improved methods for generating these models.

Objective: This study aimed to summarize the complex graphical models of MCC interactions to improve comprehension and aid analysis.

Methods: We examined the emergence of 5 chronic medical conditions (ie, traumatic brain injury [TBI], posttraumatic stress disorder [PTSD], depression [Depr], substance abuse [SuAb], and back pain [BaPa]) over 5 years among 257,633 veteran patients. We developed 3 algorithms that utilize the second eigenvalue of the graph Laplacian to summarize the complex graphical models of MCC by removing less significant edges. The first algorithm learns a sparse probabilistic graphical model of MCC interactions directly from the data. The second algorithm summarizes an existing probabilistic graphical model of MCC interactions when a supporting data set is available. The third algorithm, which is a variation of the second algorithm, summarizes the existing graphical model of MCC interactions with no supporting data. Finally, we examined the coappearance of the 100 most common terms in the literature of MCC to validate the performance of the proposed model.

Results: The proposed summarization algorithms demonstrate considerable performance in extracting major connections among MCC without reducing the predictive accuracy of the resulting graphical models. For the model learned directly from the data, the area under the curve (AUC) performance for predicting TBI, PTSD, BaPa, SuAb, and Depr, respectively, during the next 4 years is as follows—year 2: 79.91%, 84.04%, 78.83%, 82.50%, and 81.47%; year 3: 76.23%, 80.61%, 73.51%, 79.84%, and 77.13%; year 4: 72.38%, 78.22%, 72.96%, 77.92%, and 72.65%; and year 5: 69.51%, 76.15%, 73.04%, 76.72%, and 69.99%, respectively. This demonstrates an overall 12.07% increase in the cumulative sum of AUC in comparison with the classic multilevel temporal Bayesian network.

Conclusions: Using graph summarization can improve the interpretability and the predictive power of the complex graphical models of MCC.

(JMIR Med Inform 2020;8(6):e16372) doi:10.2196/16372

KEYWORDS

graphical models; graph summarization; graph Laplacian; disease network; multiple chronic conditions

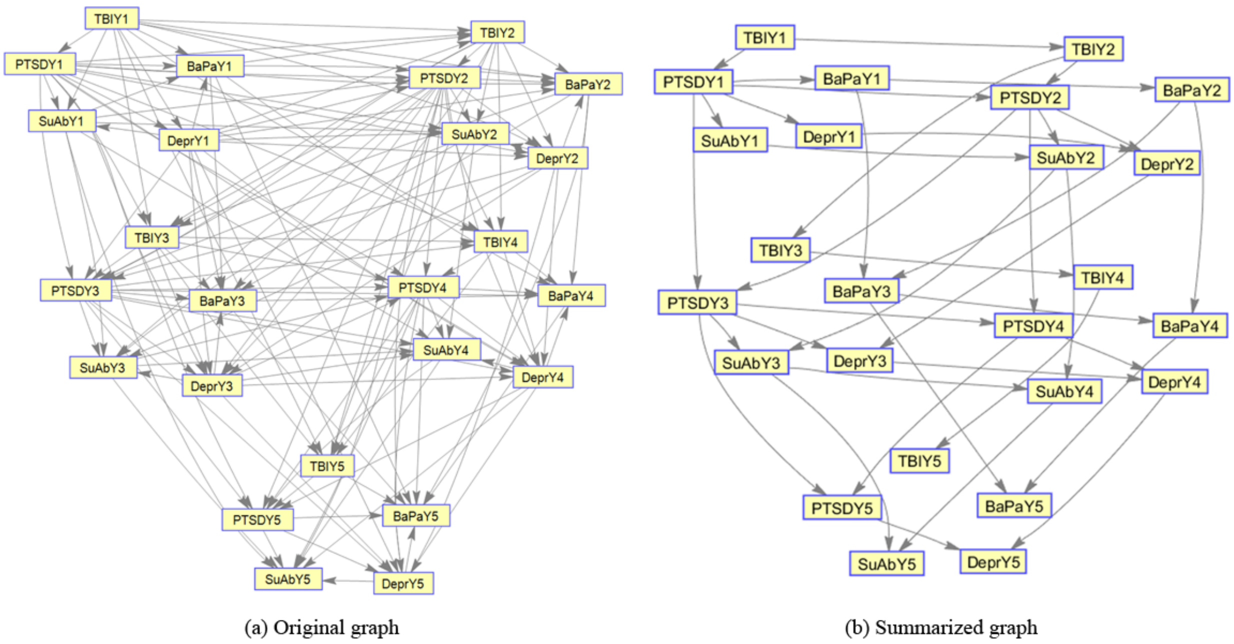
Introduction

Background

Clinical data on multiple chronic conditions (MCC) are often complex [1-4] and large [5-8]. These challenging data sets can be effectively represented in terms of graphical models [4,9]. A graphical model expresses the conditional dependencies among variables (MCC) using graph structures, where the dependencies are represented by directed or undirected edges and the variables are represented by nodes [10,11]. Analyzing these graph structures enables us to get an insight into the interactions among different chronic conditions as well as the path toward developing MCC [12]. Graphical models can also be used for the (quantitative) prediction of the occurrence versus nonoccurrence of new chronic conditions over time, based on the existing conditions, sociodemographic factors, and so on [4,13-15]. With the advancement of medical technology, the amount of data collected from different electronic medical records systems is increasing. Thus, such disease interaction graphs are becoming larger and more complex. For example, a graphical model to characterize the interaction among 30 MCC over time requires more than 1 billion edges to investigate, or a temporal graphical model to represent the relationship among 5 MCC over 5 years (time stages) requires over 400 edges to explore. There are also numerous examples of complex networks in gene expression and molecular analysis [8,16,17]. However, a large graph may have less significant edges or noisy connections, which will affect the accuracy of analysis and slow down the learning and prediction process in big data settings. Such an unsummarized graph is shown in Figure 1 (and Figure 2). Meanwhile, medical practitioners often need more concise representation to interpret the results, such as understanding the major evolution paths of MCC for planning proper intervention [9,18].

Thus, instead of using a fully/densely connected network for analysis, choosing a network with fewer but more informative connections can improve the training and querying process. However, the main questions are as follows: (1) What are the least/most informative parts of the graphical models? (2) How can such information be leveraged to summarize graphical models without losing considerable predictive/inference accuracy? and (3) How can an algorithm of this type be applied to learn a compact graph directly from the data? Effective summarization algorithms are the ones that preserve the most important structures of the original graphical model, focus on major patterns/aspects of the data, and maintain the original graph distribution (the conditional probability distribution of the original graph). They should also be capable of querying or identifying substructures/patterns in a specific set of nodes/triads (local queries) of the graph structures as well as the complete graph (global queries) to study the global influence of conditioned states.

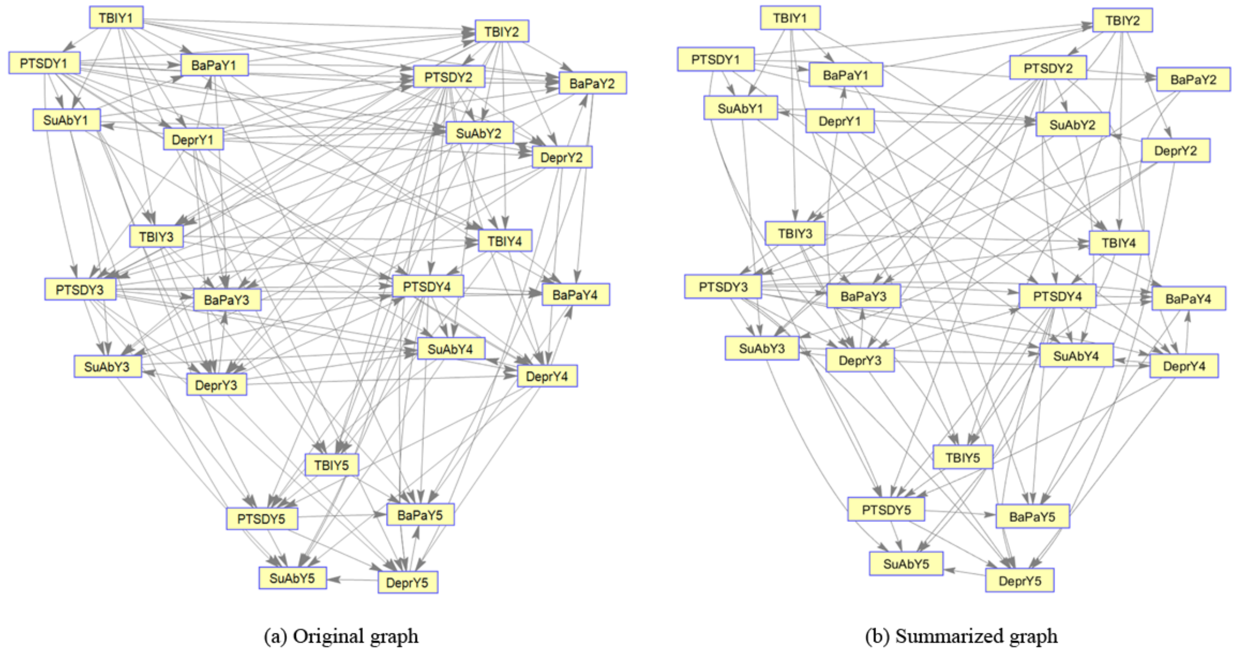
Figure 1. Learning sparse graphical models directly from emergence data on multiple chronic conditions using (a) the unsummarized graphical model (λ=0) and (b) the summarized graphical model using the EAGL structure learning algorithm (λ=1000) in which each node is a binary (0,1) variable representing the status (presence or absence) of a chronic condition in a particular year, that is, TBIY1 denotes the status of traumatic brain injury at year-1 (base year) and BaPaY5 denotes the status of back pain in year-5. BaPa: back pain; TBI: traumatic brain injury; PTSD: posttraumatic stress disorder; SuAb: substance abuse; MCC: multiple chronic conditions; EAGL: eigenvalue analysis of the graph Laplacian.



(a) Original graph

(b) Summarized graph

Figure 2. (a) Unsummarized probabilistic graphical model of the emergence of MCCs. (b) The summarized probabilistic graphical model using the EAGL summarization algorithm at a 20% summarization rate. Each node is a binary (0,1) variable representing the status (presence or absence) of a chronic condition in a particular year, that is, TBIY1 denotes the status of TBI at year-1 (base year), and BaPaY5 denotes the status of BaPa in year-5. BaPa, back pain. In the figure, BaPa: back pain; TBI: traumatic brain injury, PTSD: posttraumatic stress disorder and SuAb: substance abuse, MCC: multiple chronic condition, EAGL: eigen analysis of graph Laplacian.



(a) Original graph

(b) Summarized graph

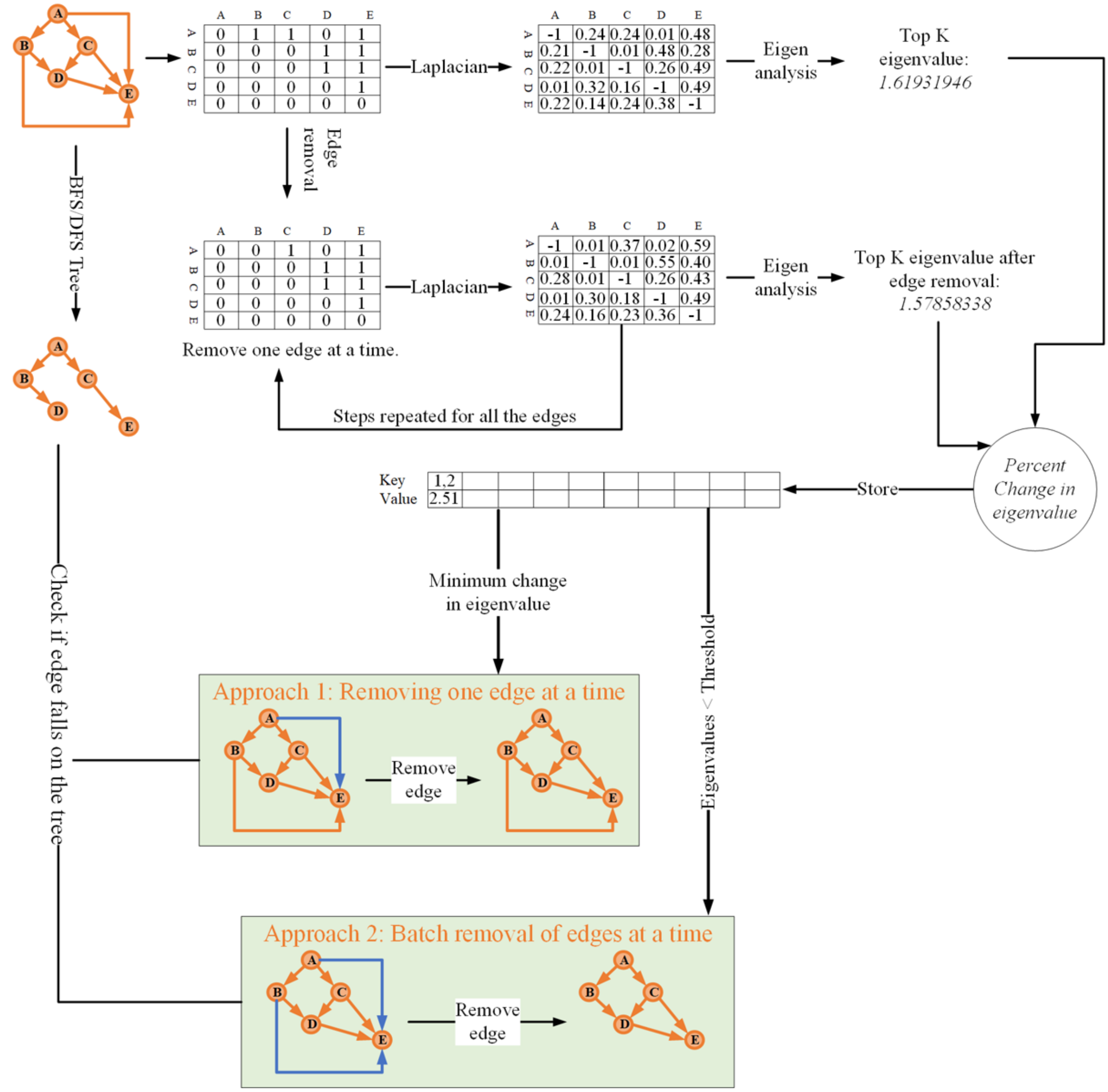
Graph summarization is also affected by factors such as data volume and complexity (structure, heterogeneity, and abstraction), dynamic/static nature of the graph, efficiency of the inference procedure, and computational complexity of the summarization approach [19]. Existing graph summarization approaches can be divided into 5 major categories:

1. Clustering-based approaches, which aggregate nodes into super-nodes and connect them using super-edges, including spectral clustering [20-22], coclustering [23], cross association [24], shingle ordering [25,26], GraSS [27], and COARSENET [28].
2. Community-based approaches, which aggregate all the nodes that belong to the same community and superimpose edge weights by summing up the weights of the original edges [29-32].
3. Simplification-based approaches, which remove less important nodes/edges, including OntoVis [33], EgoCentric [34], and MDL-based approaches [35-38].
4. Pattern set mining approaches, which create subgraphs based on the extracted patterns, including VNM [39], SUBDUE [40], VoG [35], Oddball [41], and Pegasus [42].
5. Node/edge immunization/deletion approaches, which select the best flow of the information from the source to the destination node, including MIOBI [43] and NetMelt [44].

Objective

In this work, we propose a graph summarization approach that utilizes the second eigenvalue analysis of the graph Laplacian (EAGL) to identify and prune less informative edges of the complex graphical models of MCC interactions. The intuition behind the proposed EAGL criterion is that the eigenvalue of the graph Laplacian of a graphical model is an effective measure of the connectivity and information flow [45,46]. The eigenvalue of the graph Laplacian also captures graph robustness, clustering coefficient, node importance, and several other properties [47,48]. The proposed simplification method can be utilized to (1) learn a sparse graphical model of MCC interactions directly from the data by adding a regularization term to an existing score-based structure learning algorithm to achieve a desired level of sparsity or (2) summarize a given graph of MCC interactions by removing less significant edges (with or without supporting data set) to speed up the inference process without sacrificing the predictive accuracy considerably (Figure 3). We applied the proposed approach to study conditional relationships (dependencies) among 5 multiple chronic medical conditions, including posttraumatic stress disorder (PTSD), traumatic brain injury (TBI), depression (Depr), back pain (BaPa), and substance abuse (SuAb), as well as most commonly related (coappeared) terms in the literature of MCC.

Figure 3. Visual representation of the proposed EAGL Algorithm for summarizing a directed probabilistic graphical model based on an available dataset.



Methods

Probabilistic Graphical Models

A probabilistic graphical model is specified as a tuple, $B = (G,P)$, where G denotes a graph that may be directed acyclic (in Bayesian networks, BN) or undirected (in a Markov random field), and $P(X_1,X_2,\ldots\ \ldots,X_k)$ denotes the joint probability distribution defined by conditional probabilities of the form $P(X = x_k|Pa(X = x_{k-1}))$, where X (upper case) denotes the conditional variables, x (lower case) denotes the associated values of the conditional variables, and $Pa(X = x_{k-1})$ denotes the parents of a X [9-11,49-52]. G (V,E) consists of vertices (V), that is, MCC conditions, and arcs/edges (E), that is, MCC interactions/connections, corresponding to the random variables of consideration. The network represents the joint distribution over the random variables/nodes, which can be factored according to the dependencies represented in the graph, resulting in the decomposition property of the BN:

The decomposition property makes the Bayesian inference process simple. This model is also known as the recursive model. Here, we use binary variables (nodes) representing having or not having a chronic condition (TBI, PTSD, BaPa, Depr, and SuAb) for the probabilistic graphical models.

Graph Laplacian

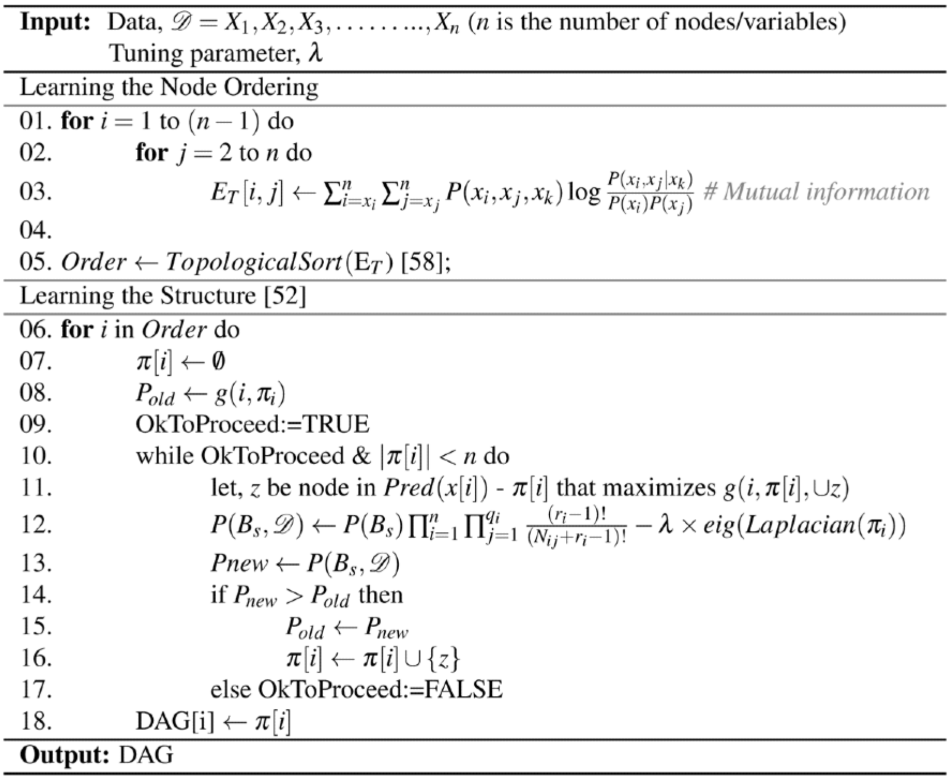
The graph Laplacian is a matrix representation of a graph, which can be used to study various properties of a graph. The first and second smallest eigenvalue of the graph Laplacian can be used to extract useful information such as graph communities (first smallest eigenvalue) and sparsest cut in a graph (second smallest eigenvalue) [45,53]. For an undirected graph, *G* (*V*,*E*), the graph Laplacian *L*(*G*) is defined as *L* = *D*–*A*, where *A* is the adjacency matrix, *D* is the degree matrix, and the elements of *L* are defined as follows [45,54]:

For a directed graph, we can consider both in- and out-degree to form the degree matrix [55,56]. In this work, we used the algorithm proposed by Fan et al [56] for deriving the graph Laplacian of directed graphical models, which is one of the most prominent methods in the literature and is straightforward to implement.

Summarizing While Learning the Structure of the Probabilistic Graphical Models Directly From Data

Figure 4 presents the major steps of the proposed EAGL algorithm for learning the sparse probabilistic graphical model structure directly from the data. The algorithm utilizes an iterative score-based method (K2, min-max hill-climbing, etc) to learn the edges (relationship) between nodes [49,57] while incorporating an active learning regularization term based on the second eigenvalue of the Laplacian of the adjacency matrix (graph Laplacian) of the graph from its previous iteration to penalize for the inclusion of less informative edges. The size of the regularization term is controlled by changing the tuning parameter λ to achieve the desired level of sparsity. In this paper, we considered the *maximum weight spanning tree (MWST) + K2* algorithm as the base learning algorithm along with the second eigenvalue of the graph Laplacian to learn a sparse structure for the probabilistic graphical model from the data. For a given data set, the MWST algorithm [50] is used to learn the initial node ordering [58]. Utilizing the ordered nodes, a greedy search method such as K2 algorithm incrementally learns the directed acyclic graph (DAG) structure from the data [52]. The regularization term is added to the K2 score function to learn the sparse representation of the DAG structure. The analysis of the computation complexity of the EAGL algorithm is provided in the *Computational Complexity* subsection.

Figure 4. Algorithm for summarizing while learning the structure of the probabilistic graphical models directly from data.



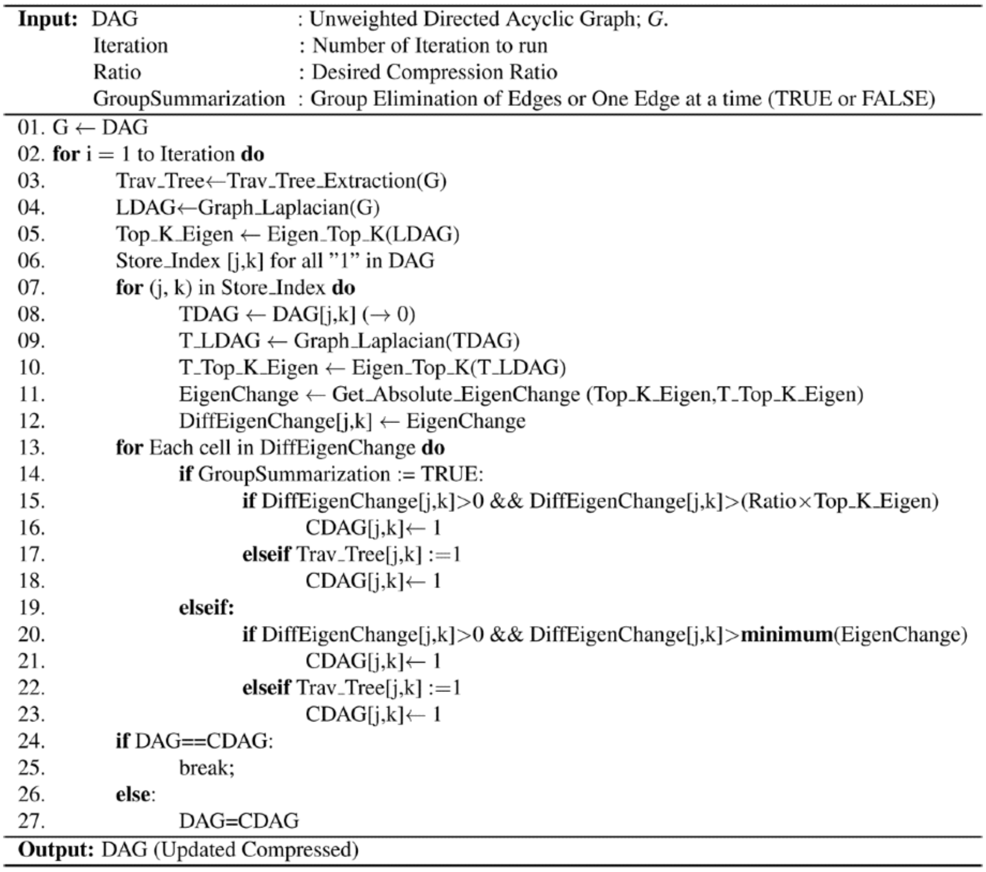
```
Input:  Data, $\mathscr{D} = X_1, X_2, X_3, ........, X_n$ ($n$ is the number of nodes/variables)
        Tuning parameter, $\lambda$
Learning the Node Ordering
for $i = 1$ to $(n-1)$ do
    for $j = 2$ to $n$ do
        $E_T[i,j] \leftarrow \sum_{i=x_i}^{n} \sum_{j=x_j}^{n} P(x_i,x_j,x_k) \log \frac{P(x_i,x_j|x_k)}{P(x_i)P(x_j)}$ # Mutual information

$Order \leftarrow TopologicalSort(\mathrm{E}_T)$ [58];
Learning the Structure [52]
for $i$ in $Order$ do
    $\pi[i] \leftarrow \emptyset$
    $P_{old} \leftarrow g(i, \pi_i)$
    OkToProceed:=TRUE
    while OkToProceed & $|\pi[i]| < n$ do
        let, $z$ be node in $Pred(x[i])$ - $\pi[i]$ that maximizes $g(i, \pi[i], \cup z)$
        $P(B_s, \mathscr{D}) \leftarrow P(B_s) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} - \lambda \times eig(Laplacian(\pi_i))$
        $Pnew \leftarrow P(B_s, \mathscr{D})$
        if $P_{new} > P_{old}$ then
            $P_{old} \leftarrow P_{new}$
            $\pi[i] \leftarrow \pi[i] \cup \{z\}$
        else OkToProceed:=FALSE
    DAG[i] $\leftarrow \pi[i]$
Output: DAG
```

Summarizing an Existing Probabilistic Graphical Model With Supporting Data

Figure 5 presents the major steps of the proposed EAGL algorithm for summarizing probabilistic graphical models when a supporting data set is available. The algorithm starts with a given probabilistic graphical model and drops edges one at a time while monitoring the changes in the second eigenvalue of the graph Laplacian. Then, it prunes the edge/s with minimum changes (removal) in the second eigenvalue of the graph Laplacian. There are 2 possible strategies for pruning the edges: (1) single edge removal—where at each stage it prunes the edge with the minimum change in the second eigenvalue—and (2) multiple edge removal—where at each stage it prunes all the edges whose change in second eigenvalue is less than a preset value (eg, 0.05). The algorithm then stops when further pruning the remaining edges change will result in a significant change in the second eigenvalue (ie, >0.05). Once all the noninformative edges have been pruned, the conditional dependencies are updated based on the supporting data. The analysis of the computation complexity of the algorithm is provided in the *Computational Complexity* subsection.

Figure 5. Algorithm for summarizing an existing probabilistic graphical model with supporting data.



```
Input: DAG                  : Unweighted Directed Acyclic Graph; G.
       Iteration            : Number of Iteration to run
       Ratio                : Desired Compression Ratio
       GroupSummarization   : Group Elimination of Edges or One Edge at a time (TRUE or FALSE)

G ← DAG
for i = 1 to Iteration do
    Trav_Tree←Trav_Tree_Extraction(G)
    LDAG←Graph_Laplacian(G)
    Top_K_Eigen ← Eigen_Top_K(LDAG)
    Store_Index [j,k] for all "1" in DAG
    for (j, k) in Store_Index do
        TDAG ← DAG[j,k] (→ 0)
        T_LDAG ← Graph_Laplacian(TDAG)
        T_Top_K_Eigen ← Eigen_Top_K(T_LDAG)
        EigenChange ← Get_Absolute_EigenChange (Top_K_Eigen,T_Top_K_Eigen)
        DiffEigenChange[j,k] ← EigenChange
    for Each cell in DiffEigenChange do
        if GroupSummarization := TRUE:
            if DiffEigenChange[j,k]>0 && DiffEigenChange[j,k]>(Ratio×Top_K_Eigen)
                CDAG[j,k]← 1
            elseif Trav_Tree[j,k] :=1
                CDAG[j,k]← 1
        elseif:
            if DiffEigenChange[j,k]>0 && DiffEigenChange[j,k]>minimum(EigenChange)
                CDAG[j,k]← 1
            elseif Trav_Tree[j,k] :=1
                CDAG[j,k]← 1
    if DAG==CDAG:
        break;
    else:
        DAG=CDAG
Output: DAG (Updated Compressed)
```

Summarizing an Existing Graphical Model Without Supporting Data

Excluding the step/s to update the remaining conditional dependencies in Figure 5 (after dropping each edge) will result in the summarization algorithm with no supporting data (see the subsection *Summarizing a Graphical Model of Multiple Chronic Conditions Terms With No Supporting Data* for results).

Structural Constraints

To avoid creating isolated nodes or islands (cluster of isolated nodes) that affect the accuracy of inference and prediction (especially in temporal graphical models), we use graph traversal methods, specifically depth-first search (DFS) [59] to preserve a path between the root and leaf nodes (for information passing between nodes). The path attained from the graph traversal is considered as a constraint in the EAGL algorithm.

Dynamic Graph

Considering the consecutive time instances of the dynamic graph, that is, t and t + 1, as a static graph, and applying appropriate structural constraints as discussed above, that is, DFS, the EAGL algorithm can be used to summarize dynamic graphical models as well.

Results

Study Population

The relationship among the emergence of MCC can be expressed effectively using probabilistic graphical models, where nodes represent the emergence of chronic conditions, that is, BaPa, Depr, and so on, and edges show the statistical relationship (conditional dependency) between them (BaPa and Depr). Here, we are interested in sparse learning of the structure and parameters of the probabilistic graphical model using the EAGL algorithm based on an available data set of the emergence of MCC. Our deidentified data were collected from a large national cohort of US military veteran patients (N=608,503), who were deployed in support of the wars in Afghanistan and Iraq and began receiving care in the Veterans Health Administration (VA) between 2002 and 2011. For the purpose of this analysis, we have only considered patients who received care each year for the first 5 years after entering VA care (N=257,633). Dropout may result from not requiring care, dropping out of VA care, or death. This study received institutional review board approval from the University of Texas Health Science Center at San Antonio and the Bedford VA Hospital, with a waiver of informed consent. A summary of the study population is shown in Table 1.

Table 1. Demographics of the patients included in the study.

Demographics	Serial number					
	1	2	3	4	5	6
Race	White	Black	Hispanic	Asian	Native	Unknown
Gender, n (%)						
Male	148,355 (57.58)	35,758 (13.88)	25,373 (9.85)	5639 (2.19)	3081 (1.20)	2135 (0.83)
Female	19,183 (7.45)	11,828 (4.59)	4232 (1.64)	981 (0.38)	707 (0.27)	361 (0.14)
Marital status, n (%)						
Married	74,487 (28.91)	23,308 (9.05)	14,523 (5.64)	3067 (1.19)	1747 (0.68)	1346 (0.52)
Unmarried	93,051 (36.12)	24,278 (9.42)	15,082 (5.85)	3553 (1.38)	2041 (0.79)	1150 (0.45)
Age group (years), n (%)						
18-30	96,799 (37.57)	20,047 (7.78)	17,016 (6.60)	3235 (1.26)	2115 (0.82)	1062 (0.41)
31-40	36,003 (13.97)	12,468 (4.84)	6606 (2.56)	1361 (0.53)	925 (0.36)	625 (0.24)
41-50	26,167 (10.16)	12,710 (4.93)	4758 (1.85)	1564 (0.61)	564 (0.22)	673 (0.26)
≥51	8569 (3.33)	2361 (0.92)	1225 (0.48)	460 (0.18)	184 (0.07)	136 (0.05)
Education, n (%)						
Unknown	2334 (0.91)	658 (0.26)	386 (0.15)	131 (0.05)	60 (0.02)	51 (0.02)
Less than high school	2037 (0.79)	504 (0.20)	360 (0.14)	60 (0.02)	60 (0.02)	22 (0.01)
High school graduate	129,921 (50.43)	37,506 (14.56)	23,592 (9.16)	4732 (1.84)	3004 (1.17)	1808 (0.70)
Some college	16,743 (6.50)	4819 (1.87)	2933 (1.14)	598 (0.23)	376 (0.15)	287 (0.11)
College graduate	12,024 (4.67)	3160 (1.23)	1893 (0.73)	879 (0.34)	217 (0.08)	223 (0.09)
Post college education	4479 (1.74)	939 (0.36)	441 (0.17)	220 (0.09)	71 (0.03)	105 (0.04)

Learning Sparse Probabilistic Graphical Models Directly From Data

The EAGL algorithm begins with a DAG structure provided by a score-based algorithm [9,49], that is, MWST + K2. It then calculates the second eigenvalue of the graph Laplacian for the obtained DAG. Next, it multiplies the second eigenvalue with a tuning parameter. It adds it as a penalty term to the main scoring function to determine which edges to remove for the next iteration. The last 2 steps are repeated until a stopping criterion is met.

Figure 1 illustrates 2 graphical models, which have been estimated with different choices of the tuning parameter (λ) to control the sparsity in the EAGL algorithm: (1) the unsummarized graphical model without a penalty (λ=0) and (2) a summarized graphical model with a large tuning parameter (λ=1000). The tuning parameter was set at λ=0 (Figure 4), which results in an unsummarized graphical model [9] that provides a year 2 predictive accuracy of TBI=75.69%, PTSD=78.97%, BaPa=63.16%, SuAb=72.93%, and Depr=68.24%, compared with 72.34% reduction in the number of edges, and year 2 predictive accuracy of TBI=79.91%, PTSD=84.04%, BaPa=78.83%, SuAb=82.50%, and Depr=81.47% for the summarized graphical model (λ=1000; Figure 1, summarized graph; Table 2).

To evaluate the model, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve [60] was considered. ROC curves are tools used to illustrate the diagnostic ability of a binary classifier at different threshold values. The curves are created by plotting the true positive rate (probability of detection) against the false positive rate (false detection ratio) at the threshold settings. This plot can be summarized into a single metric by calculating the area under the ROC curve. The AUC identifies how much a model is capable of distinguishing between different classes. AUC values range between 0 and 1, with higher values representing better classification accuracy. Table 2 illustrates the predictive accuracy of the learned graphical model under different choices of tuning parameters $\lambda=0,10^{-2},10^{-1},...,10^{5}$ (λ=0 represents the classical/unsummarized graphical model) using the AUC metrics based on 10-fold cross-validation. It also shows the predictive performance of the learned graphical model using the popular Akaike information criterion (AIC). The superior predictive accuracy of the sparse graphical model by the EAGL algorithm can be attributed to the removal of spurious (less significant edges) edges in the graph, which improves the information propagation through high-confidence paths on the graph. Table 2 also compares the performance of the EAGL with another popular approach, AIC, which achieves 66.67% edge removal and year 2 predictive accuracy of TBI=59.49%, PTSD=63.45%, BaPa=78.51%, SuAb=61.32%, and Depr=59.05%.

Table 2. The area under the curve performance of the sparse probabilistic graphical model learned by the eigenvalue analysis of the graph Laplacian algorithm directly from the data with different choices of tuning parameters ($\lambda=0,10^{-2},10^{-1},...,10^{5}$) for predicting future comorbidities (year 2 to year 5), given the comorbidity information of the past year (year 1), along with the area under the curve performance of a comparing algorithm, namely, Akaike information criterion (AIC) as well as the associated summarization ratios.

Prediction year	Lambda									
	0.00	0.01	0.10	1.00	10.00	100.00	1000.00	10000.00	100000.00	AIC
Year 2 (%)										
TBI[a]	75.69	75.69	75.69	75.88	76.69	79.57	79.91	79.88	79.88	59.49
PTSD[b]	78.97	78.97	79.08	79.53	81.31	83.11	84.04	83.70	83.70	63.45
BaPa[c]	63.16	63.16	63.16	63.63	48.57	78.29	78.83	78.82	78.82	78.51
SuAb[d]	72.93	72.93	73.04	73.33	75.22	74.58	82.50	85.00	85.00	61.32
Depr[e]	68.24	68.24	68.27	68.45	71.02	74.26	81.47	81.61	81.61	59.05
Year 3 (%)										
TBI	72.22	72.22	72.19	72.63	74.82	76.28	76.23	76.11	76.11	62.28
PTSD	76.01	76.01	76.02	76.84	78.71	80.11	80.61	80.35	80.35	61.95
BaPa	60.92	60.92	60.98	61.82	70.07	73.15	73.51	73.84	73.84	73.27
SuAb	70.80	70.80	70.82	71.02	73.62	68.51	79.84	81.13	81.13	61.83
Depr	65.16	65.16	65.18	65.93	69.01	70.48	77.13	77.10	77.10	56.09
Year 4 (%)										
TBI	70.86	70.86	70.81	71.11	72.96	73.20	72.38	72.39	72.39	60.71
PTSD	73.21	73.21	73.35	74.11	75.97	78.00	78.22	77.84	77.84	61.97
BaPa	61.81	61.81	61.82	62.50	69.96	72.98	72.96	72.61	72.61	72.84
SuAb	68.97	68.97	68.82	68.34	70.88	74.96	77.92	79.64	79.64	60.72
Depr	64.73	64.73	64.79	65.09	67.29	68.00	72.65	73.54	73.54	56.23
Year 5 (%)										
TBI	70.88	70.88	70.92	71.78	72.50	73.47	69.51	69.38	69.38	59.70
PTSD	72.72	72.72	72.88	73.43	75.21	76.50	76.15	74.86	74.86	60.59
BaPa	53.46	53.46	53.41	54.09	69.63	72.64	73.04	68.52	68.52	72.30
SuAb	63.73	63.73	63.74	61.34	63.46	73.65	76.72	77.26	77.26	61.46
Depr	64.01	64.01	64.11	64.87	66.58	67.89	69.99	71.37	71.37	56.07
Edge details										
Edges, n	141	140	139	128	107	73	39	24	24	47
Edge removal (%)	0.00	0.71	1.42	9.22	24.11	48.23	72.34	82.98	82.98	66.67

[a]TBI: traumatic brain injury.

[b]PTSD: posttraumatic stress disorder.

[c]BaPa: back pain.

[d]SuAb: substance abuse.

[e]Depr: depression.

Figure 6 studies the relationship between the changes in the tuning parameters and the second eigenvalue of the graph Laplacian, which shows no change (in the second eigenvalue) over very small/large choices of the tuning parameters and logarithmic growth over other (midrange) choices of the tuning parameter. From Figure 7, we observed a similar pattern between changes in the tuning parameters and model sparsity and predictive accuracy, where very small (<0.01) or very large (>10^{4}) changes in the tuning parameter did not improve the edge removal rate and/or predictive accuracy. Meanwhile, other choices of tuning parameters generally improve both sparsity and predictive accuracy. Therefore, the change in the second eigenvalue of the graph Laplacian can be used as a stopping criterion for EAGL algorithm; specifically, when increasing the tuning parameter does not change the second eigenvalue of the graph Laplacian, the algorithm shall stop (the analysis of first eigenvalue is provided in Multimedia Appendix 1).

Figure 6. The relationship between the change in the tuning parameter (λ) and the second eigenvalue.

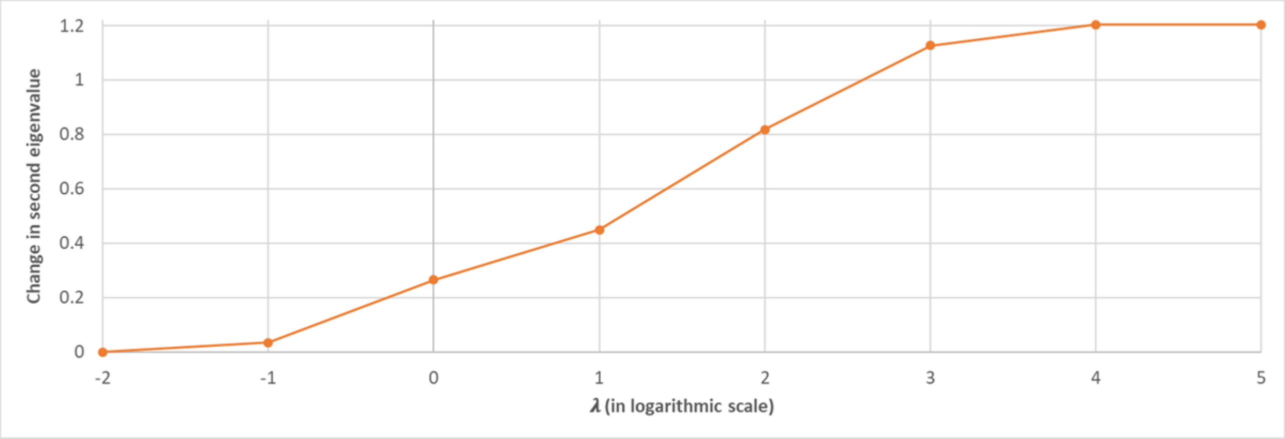
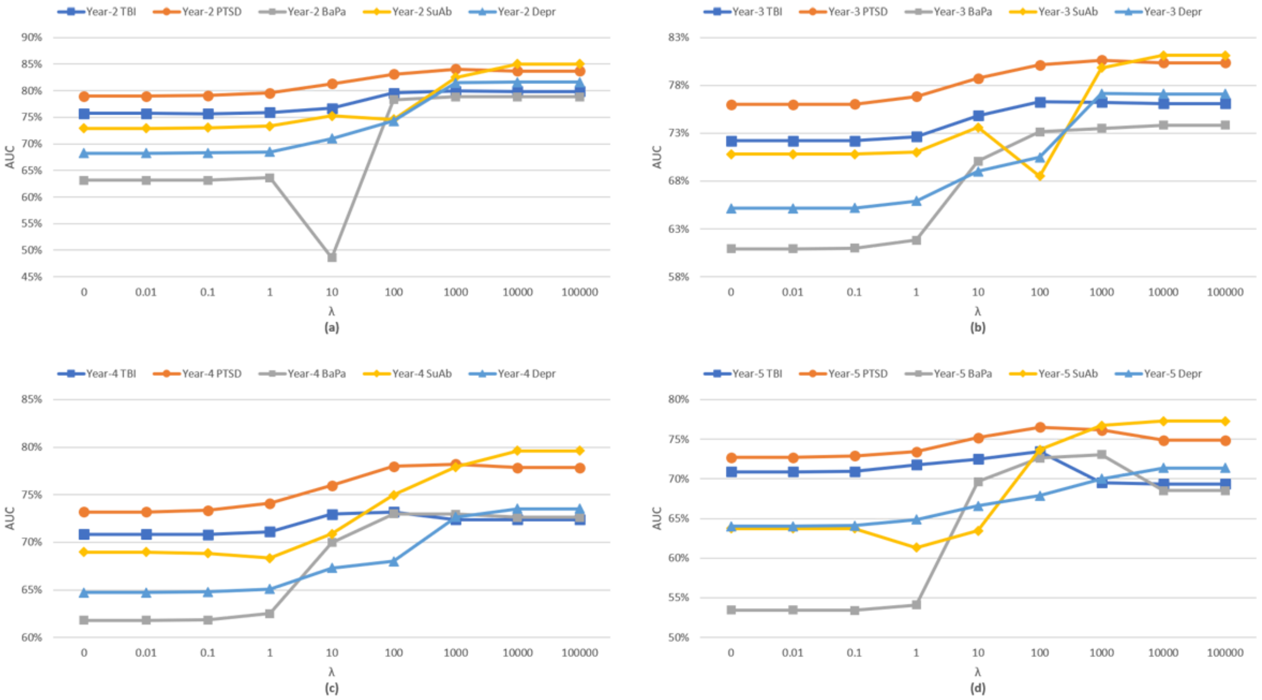


Figure 7. The relationship between the change in the tuning parameters (λ) and the area under the curve: (a) year-2; (b) year-3; (c) year-4; (d) year-5 of the study. BaPa: back pain; TBI: traumatic brain injury, PTSD: posttraumatic stress disorder and SuAb: substance abuse, MCC: multiple chronic conditions, EAGL: eigen analysis of graph Laplacian.



Summarizing an Existing Probabilistic Graphical Model With Supporting Data

In many real-life situations, we are given a graphical model that could potentially be simplified. The EAGL algorithm, which is based on the second eigenvalue of the graph Laplacian, can be used to identify and prune insignificant edges of the graph to achieve the desired level of summarization. The EAGL algorithm begins by calculating the second eigenvalue of the graph Laplacian of the given graphical model. It then extracts the DFS tree to determine the edges to avoid isolated nodes. Next, from the set of edges that is not lying on the DFS tree, the algorithm (temporarily) removes edges one at a time and calculates the percentage of the change in the second eigenvalue of the remaining graph Laplacian. Subsequently, it (permanently) removes the edge, resulting in a minimum change in the second eigenvalue of the graph Laplacian. The last 2 steps are repeated until a stopping criterion is met. Once the summarized network structure is attained, the weight S of the edges (conditional probabilities) are estimated using a standard parameter estimation algorithm [10,50]. Figure 3 provides a visual representation of the proposed algorithm. An example of this step-by-step process is provided in Multimedia Appendix 2.

Here, we are interested in summarizing an existing probabilistic graphical model of MCC relationships attained using a score-based method [9] based on the MCC data set discussed above (Figure 2, original graph). The summarized graph in Figure 2 illustrates the structure of the summarized graphical model based on removing less significant edges/paths of the original graphical model using the EAGL algorithm at a 20% summarization rate (removing 20% of existing edges).

Table 3 presents the AUC performance of the summarized graphical models at different summarization ratios of 0%, 1%, 5%, 10%, and 20% (0% represents the classical/unsummarized graphical model) for predicting future comorbidities (year 2 to year 5), given the year 1 comorbidity using 10-fold cross-validation. It also shows the predictive performance of the learned graphical model using the MIOBI [43] algorithm and the CHEETAH [61] algorithm at different summarization ratios. As shown in the table, the proposed EAGL algorithm generally provides the most competitive predictive accuracy among the comparing methods across different summarization ratios. This is while the EAGL algorithm also prevents the creations of island nodes, which helps with the interpretation of the results.

Although increasing the summarization ratio generally results in a sparser graphical model, for mild summarization ratios (<10%), using EAGL can also improve the predictive performance of the graphical model by preserving more informative edges/paths as it should. However, a large choice of summarization ratios (>10%) can decrease the predictive performance, depending on the topological location of the node (chronic conditions) and the associated edges that have been pruned (Table 3).

Table 3. The area under the curve performance of the original and summarized probabilistic graphical models at different summarization ratios (1%, 5%, 10%, and 20%) for predicting future comorbidities (year 2 to year 5), given the comorbidity information of the past year (year 1).

Prediction year	EAGL[a]					MIOBI [33]					CHEETAH [62]				
	Original	1.00	5.00	10.00	20.00	Original	1.00	5.00	10.00	20.00	Original	1.00	5.00	10.00	20.00
Year 2 (%)															
TBI[b]	75.69	75.63	75.53	75.09	63.34	75.69	75.78	75.63	63.99	56.08	75.69	75.70	65.25	63.95	61.80
PTSD[c]	78.97	78.94	80.20	80.87	81.51	78.97	79.32	80.57	70.86	71.04	78.97	79.12	79.54	80.19	71.15
BaPa[d]	63.16	63.18	61.05	61.37	65.53	63.16	62.99	63.14	62.44	63.25	63.16	63.20	63.43	64.16	64.80
SuAb[e]	72.93	72.95	75.74	70.59	68.26	72.93	72.93	72.88	73.54	69.99	72.93	73.10	73.78	73.96	74.29
Depr[f]	68.24	68.23	70.48	62.88	59.27	68.24	68.24	68.03	66.20	55.50	68.24	68.36	68.51	68.74	70.22
Year 3 (%)															
TBI	72.22	72.25	72.15	72.13	71.78	72.22	72.24	72.36	70.20	59.82	72.22	72.22	64.62	63.54	61.33
PTSD	76.01	75.97	77.37	77.81	78.57	76.01	76.40	77.89	75.34	76.60	76.01	76.15	76.90	76.78	69.36
BaPa	60.92	60.98	59.32	59.49	61.88	60.92	60.96	61.12	61.34	58.44	60.92	61.04	61.29	61.49	61.80
SuAb	70.80	70.81	72.49	70.54	67.58	70.80	70.73	70.72	70.70	71.08	70.80	70.83	71.53	71.73	71.97
Depr	65.16	65.20	66.95	68.29	58.28	65.16	65.13	65.16	64.45	64.24	65.16	65.37	65.84	65.99	66.85
Year 4 (%)															
TBI	70.86	70.78	71.10	70.76	70.48	70.86	70.82	70.82	69.41	68.20	70.86	70.27	64.85	64.08	61.94
PTSD	73.21	73.18	74.29	74.95	76.02	73.21	73.52	74.99	72.92	74.51	73.21	73.34	73.88	74.11	67.64
BaPa	61.81	61.81	60.24	60.38	63.01	61.81	61.76	62.01	62.62	59.29	61.81	61.93	62.14	62.43	63.03
SuAb	68.97	69.00	70.28	69.13	68.24	68.97	68.97	68.73	69.28	70.99	68.97	69.14	69.53	69.78	70.24
Depr	64.73	64.71	65.48	65.53	61.55	64.73	64.73	64.63	63.97	64.03	64.73	64.88	65.09	65.25	65.68
Year 5 (%)															
TBI	70.88	70.91	71.55	71.63	70.74	70.88	70.82	70.91	70.53	69.14	70.88	70.52	66.02	65.25	63.35
PTSD	72.72	72.70	73.55	73.75	74.34	72.71	72.95	73.91	72.24	73.61	72.71	72.85	73.19	73.39	67.30
BaPa	53.46	53.49	51.20	50.58	50.67	53.46	53.21	53.11	53.03	49.26	53.46	53.51	53.50	53.73	53.01
SuAb	63.73	63.76	64.35	62.59	59.70	63.73	63.61	63.46	63.07	61.47	63.73	63.90	63.99	63.87	64.12
Depr	64.01	63.99	64.78	64.50	61.23	64.01	63.96	63.74	63.38	63.32	64.01	64.19	64.50	64.70	65.19

[a] EAGL: Eigenvalue analysis of the graph Laplacian.

[b] TBI: traumatic brain injury.

[c] PTSD: posttraumatic stress disorder.

[d] BaPa: back pain.

[e] SuAb: substance abuse.

[f] Depr: depression.

Figure 8 presents the relationship between the various choices of compression ratio and the changes in the second eigenvalue of the graph Laplacian. As shown in the figure, for compression ratio values of >10%, the rate of change in the second eigenvalue increases. Moreover, Figure 9 provides the predictive accuracy of the summarized graph for the 5 chronic conditions in the study at different years (year 2 to year 5), which shows a reduction in the AUC for larger choices of summarization ratios (>10%). Therefore, a sharp increase in the changes in the second EAGL can be used as a stopping criterion for EAGL. (The analysis of the first eigenvalue is provided in Multimedia Appendix 1.)

Figure 8. Decrease in the second eigenvalue with reduction in the number of edges.

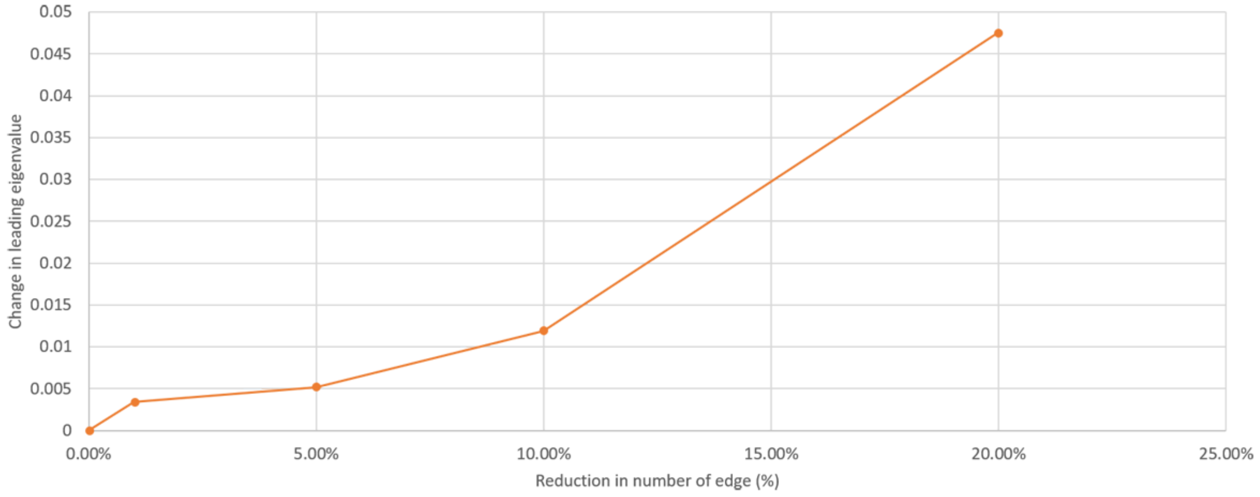
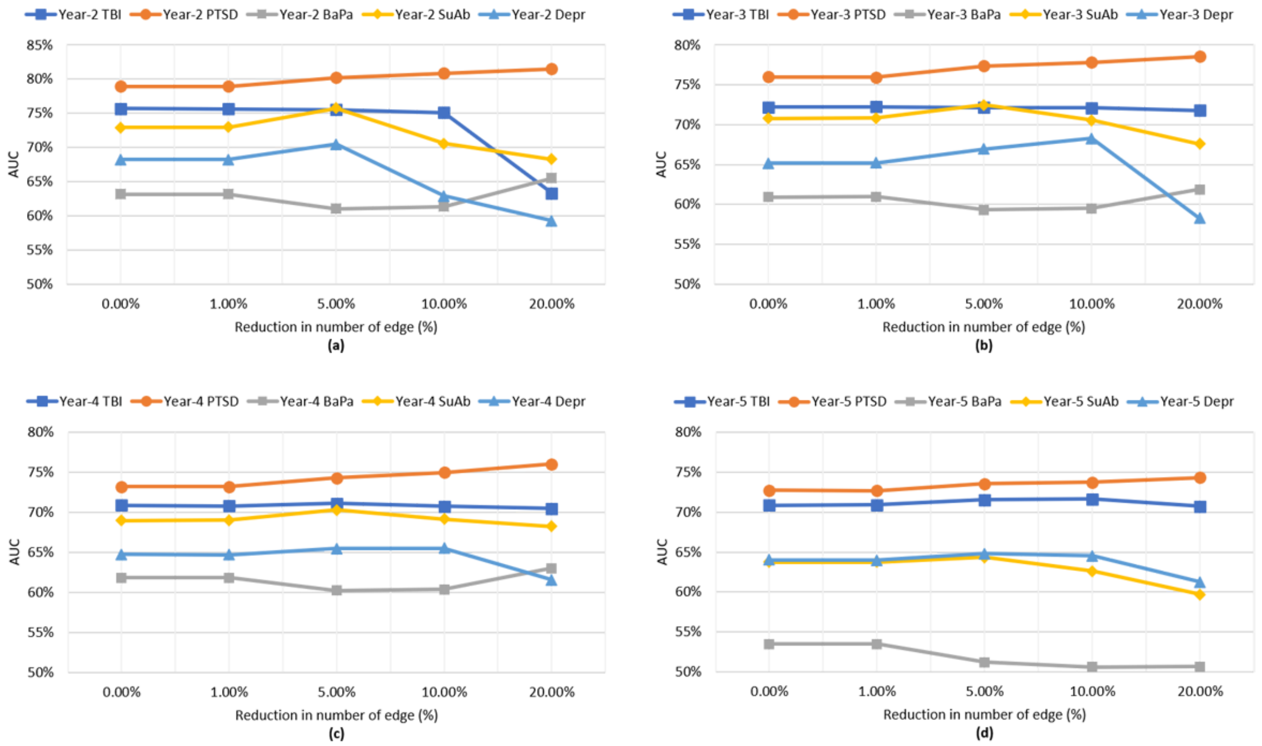


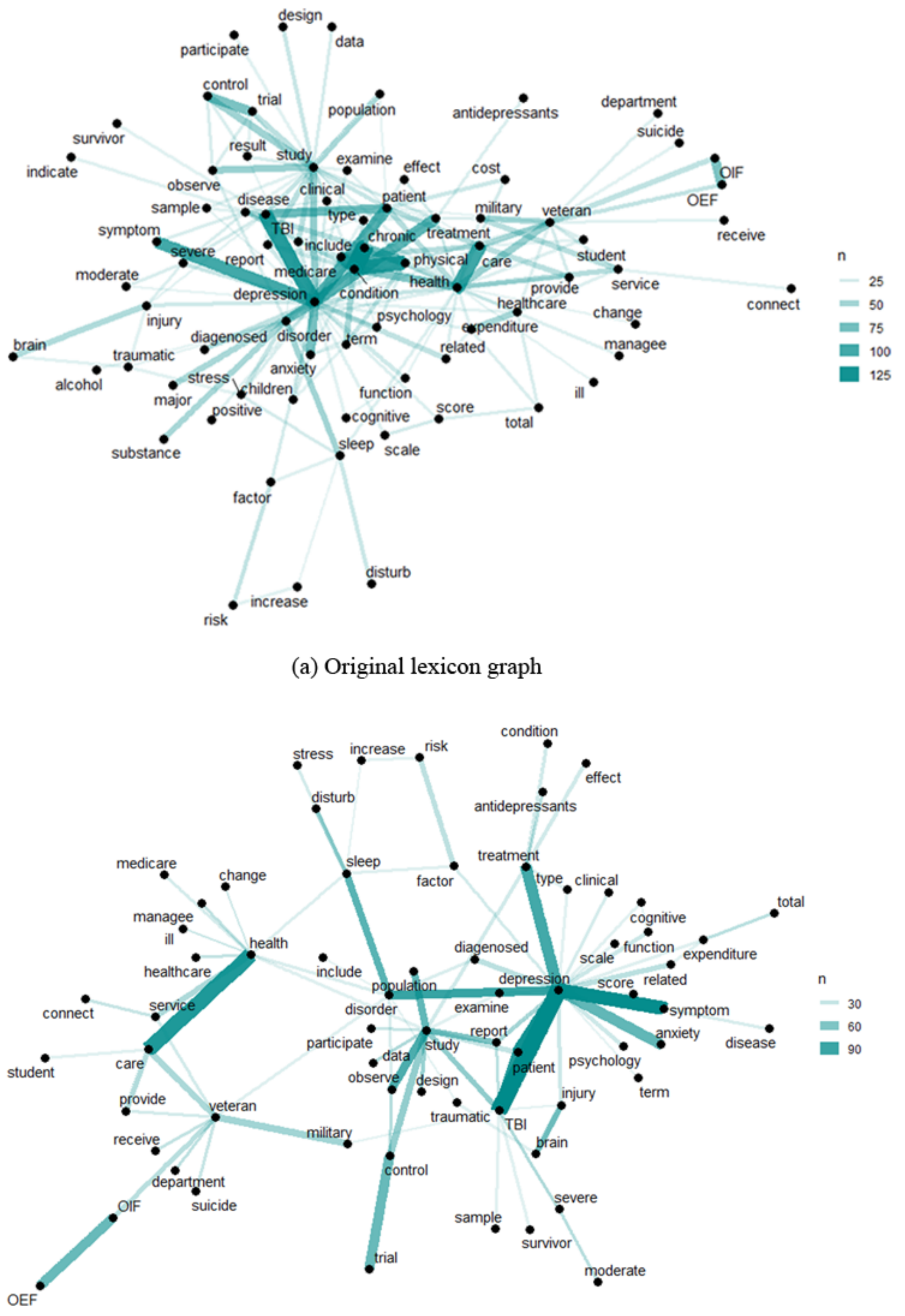
Figure 9. The relationship between the changes in the tuning parameters (λ) and the area under the curve in the second eigenvalue (λ) over (a) year-2, (b) year-3, (c) year-4, and (d) year-5 of the study. BaPa: back pain; TBI: traumatic brain injury; PTSD: posttraumatic stress disorder; SuAb: substance abuse; MCC: multiple chronic conditions; EAGL: eigenvalue analysis of the graph Laplacian.



Summarizing a Graphical Model of Multiple Chronic Conditions Terms With No Supporting Data

A lexicon graph contains a list of stems and affixes, together with basic information about them in the form of a graphical model. This is generally used to represent interconnected word pairs and their frequencies in natural language processing. Here, we are interested in exploring the opportunity to summarize a graphical model of MCC-related terms (Lexicon graph) with no supporting data using the EAGL algorithm. The graphical model was developed based on a lexicon graph from a collection of medical journals. The journals were extracted using the following keywords: Veterans, Traumatic Brain Injury, Back Pain, Post-Traumatic Stress Disorder, Depression, Substance Abuse, Chronic Diseases, Comorbidity, Multimorbidity, chronic conditions, chronic illness, and chronic pain. A total of 20 peer-reviewed journal papers were collected based on Google Scholar ranking (without expert opinion). Multimedia Appendix 3 lists the journal papers used for the creation of the lexicon graph. From the collected papers, the term and their frequencies are extracted and turned into a data set [41,62-81]. The 200 most frequent word pairs are then selected to build the lexicon graph, where the strength of the edges (connections) represents the co-occurrence of the word pairs in the same sentence (original lexicon graph in Figure 10).

Figure 10. (a) Lexicon graph of the top 200 most frequent word pairs attained from text mining of 20 medical journal papers; (b) lexicon graph after summarization algorithm (70% summarization) was performed in the graph. OEF: operation enduring freedom; OIF: operation Iraqi freedom.



(a) Original lexicon graph

(b) Summarized lexicon graph

Summarized lexicon graph in Figure 10 illustrates the summarized graphical model using the EAGL algorithm at a 70% summarization rate (edge removal) without utilizing any supporting data set. The summarized graph presents a cluster of strong relationships among chronic conditions such as <Depr, anxiety, TBI, symptoms, and treatment>. It also shows meaningful connections among <study, design, observe, population, control, and trial> and <healthcare, ill manage, service, and medicare>. There are also other interesting groups of highly connected terms such as <veteran care, military, suicide, and Operation Iraqi Freedom (OIF)> or <sleep, stress, and increased risk>. Multimedia Appendix 3 shows an enlarged version of the lexicon graph and its compressed form using the EAGL algorithm. It is worth noting that the algorithm here does not estimate/update the weight of (remaining) edges at each iteration (removal of edges); therefore, it is very efficient in summarizing large lexical graphs.

Computational Complexity

In this section, we derive the time complexity of algorithms shown in Figures 4 and 5, which is presented earlier. Let *n* denote the number of node/variables/vertices (chronic conditions), *e* denote the number of edges (relationship between pair of chronic conditions), *m* denote the number of observations/cases (patient observations), and *r* denote the number of possible values/instances for each variable (in our study *r*=2, which represents having/not having a condition). Figure 4 consists of 5 components with the following (known) computational complexities: (1) MWST for node ordering: *0* (n^2); (2) topological sorting: *0* ($n + e$); (3) graph Laplacian: *0* (n); (4) eigenvalue calculation: *0* (n^2); and (5) K2 structure learning with regularization: *0* (mn^4r). Integrating the complexities of the 5 components with some algebraic simplification, the overall complexity of Figure 4 can be derived as *0* (mn^4r).

Figure 5 also consists of 3 components with the following (known) computational complexities: (1) depth-first tree extraction: *0* ($n + e$); (2) graph Laplacian: *0* (n); and (3) eigenvalue calculation: *0* (n^2). Let *p* denote the number of edges to be removed (the desired amount of edge removal). After some algebraic operations (to account for the loops), the overall complexity of Figure 5 can be derived as *0* (en^2p)

Discussion

Principal Findings

Graphical models are increasingly being used for descriptive, predictive, and prescriptive analytics in various applications, including social media, computer networks, genetics, and disease prognosis [7,8,82-84]. The effectiveness of a graphical model depends on the quality of the information propagating through nodes, which is affected by the topology of the network. Graph topology also affects other properties of a graphical model, including complexity, robustness, and scalability [85]. A fully connected network can be considered the most robust in terms of information dissemination but may cause overfitting, slow training, and memory allocation issues. Graph summarization can be performed to identify the important structures, major patterns, and dissemination of information in complex graphical models of MCC interaction.

In this study, we have addressed the problem of summarizing complex graphical models and identifying their important patterns by modifying the edges of the graph. These types of graphical frameworks are useful for analyzing plausible interactions between disease states [4]. The eigenvalue of the graph Laplacian reveals the characteristics of a graph. For a large graph, the second eigenvalue of the graph Laplacian determines the amount of information that is being distributed by the graph. Thus, by analyzing the second eigenvalue of the graph Laplacian, we attain a measure (EAGL) of sparse cutoff. The proposed EAGL algorithm can be used as an active learning unsupervised method to directly learn a sparse probabilistic graphical model from an available data set or summarize an existing graphical model with or without a supporting data set.

The first approach (using direct learning) results in a refined model where network analysis can be performed by an end user with specific needs and expertise. Our direct learning model (Figure 4) demonstrates very good performance when data are available, and the algorithm is able to learn de novo. This results in a graph (Figure 1) with predictive abilities that can be interpreted by clinicians and medical researchers with an understanding of the medical conditions of interest.

The second approach (Figure 5) summarizes an existing graphical model with or without a supporting data set. The EAGL algorithm, which is based on a simplification-based rule edge removal strategy, can also be used to reveal important patterns within a given graphical model by removing the edges with a marginal contribution to the leading eigenvalue of the graph Laplacian.

Our findings revealed that the proposed summarization algorithm can indeed improve the predictive accuracy of the summarized graphical model while reducing its size and increasing the inference efficiency. We used 2 data sets of (1) 257,633 veteran patients who have been monitored for the emergence of 5 multiple conditions (TBI, PTSD, BaPa, Dep, and SuAb) over 5 years and (2) the coappearance of the 200 most frequent word pairs in the literature of MCC to validate the performance of the proposed EAGL approach.

Although the statistical details of the proposed model might be complex for some practitioners to understand, the resulting algorithm can be seen as a step toward creating more interpretable analytical models for understanding the evolution of MCC, by removing less informative edges in complex networks of MCC (resulting in a sparser network), without losing predictive accuracy. In fact, practitioners do not need to know the details of the proposed algorithm to utilize it. They can use a simple tuning parameter (λ) to control the level of resulting network sparsity (number of remaining edges), that is, setting a high value for the tuning parameter results in a very sparse network (with few edges), which is easy to understand (Figures 1 and 2). Such a (sparse) graphical representation provides a straightforward visualization of how the presence of one condition can affect the emergence of another condition without complex statistics. It also helps interpret the probabilistic results from statistical analysis.

Finally, the proposed EAGL approach can help medical practitioners and health care analysts not only in terms of developing a predictive tool to analyze the probability of a new chronic condition development, given the existing conditions (Figures 6-9), but also by using a tuning parameter (λ) to identify major interaction patterns among MCC. The model can also be used as a visualizing tool to inspect the interaction among MCC (Figures 1 and 2).

Limitations

Although the proposed EAGL algorithm successfully extracts important connections and controls the level of sparsity, it has a few limitations and potential problems. Algorithm presented in Figure 4 needs to be built on top of a structure learning model.

In this study, we utilized the MWST + K2 method [9]. This is a heuristic-based structure learning model, where the initial node order has to be known or learned using the MWST method. Algorithm presented in Figure 5 requires an appropriate tree extraction method to ensure that there will be no island node (or set of nodes), which can limit the level of summarization. In addition, for a high summarization ratio, the summarization algorithm can decrease the prediction accuracy. Finally, both algorithms (Figures 4 and 5) primarily target acyclic graphs, but their usefulness to depict complex webs of causation in chronic conditions, which can involve loops (particularly of reinforcing types), is limited.

Conclusions

In this work, we propose a graph summarization approach that utilizes the second eigenvalue of the graph Laplacian to identify and prune less informative edges of the complex graphical models of MCC interaction. We developed 3 algorithms based on the proposed approach to deal with different scenarios with respect to the availability of data and/or a graphical model. The first algorithm learns a sparse graphical model of MCC interactions directly from the data by regularizing an existing score-based structure learning algorithm to achieve a desired level of sparsity. The second algorithm summarizes an existing graph of MCC interactions by removing less informative connections with respect to a supporting data set. The third algorithm simplifies a given MCC graph by removing the less important edges without a supporting data set. We validated the performance of the first 2 algorithms based on a large data set of veteran patients who have been monitored for over 5 years and 5 multiple chronic medical conditions, including PTSD, TBI, Depr, BaPa, and SuAb. We also validated the third algorithm based on a data set of coappearances of the 200 most frequent word pairs in the literature of MCC. The results showed that the proposed EAGLE algorithm effectively extracts important connections and dependency patterns from the complex graphical model of the interactions of MCC. It can also control the level of sparsity in the resulting graph based on the practitioners' needs using a simple tuning parameter. Finally, it improves the predictive accuracy of the resulting summarized graphical model.

Acknowledgments

This research work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number 1SC2GM118266-01 and the US Department of Veterans Affairs funds, I01HX000329 and IK6HX002608 to MP. The sponsors played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the US Government or the US Department of Veterans Affairs, and no official endorsement should be inferred.

Authors' Contributions

SF, AA, and SS developed the EAGL algorithms. SF preprocessed the data, coded the algorithms, and conducted the numerical studies. CC prepared the terms for the MCC database and Figure 10 and Multimedia Appendix 3. CJ, AA, SS, MP, PR, and JW reviewed and analyzed the results, and SF, AA, and CJ wrote the manuscript. All authors reviewed the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results of the eigenvalue analysis of the graph Laplacian algorithm based on the first eigenvalue.
[DOCX File , 361 KB - medinform_v8i6e16372_app1.docx]

Multimedia Appendix 2
A sample example of the eigenvalue analysis of the graph Laplacian algorithm for a small graph.
[DOCX File , 117 KB - medinform_v8i6e16372_app2.docx]

Multimedia Appendix 3
Multiple chronic conditions term lexicon.
[DOCX File , 418 KB - medinform_v8i6e16372_app3.docx]

References

1. Faruqui SH, Du Y, Meka R, Alaeddini A, Li C, Shirinkam S, et al. Development of a deep learning model for dynamic forecasting of blood glucose level for type 2 diabetes mellitus: secondary analysis of a randomized controlled trial. JMIR Mhealth Uhealth 2019 Nov 1;7(11):e14452 [FREE Full text] [doi: 10.2196/14452] [Medline: 31682586]

2. Fitzpatrick SL, Hill-Briggs F. Measuring health-related problem solving among African Americans with multiple chronic conditions: application of Rasch analysis. J Behav Med 2015 Oct;38(5):787-797. [doi: 10.1007/s10865-014-9603-4] [Medline: 25319236]
3. Alaeddini A, Jaramillo CA, Faruqui SH, Pugh MJ. Mining major transitions of chronic conditions in patients with multiple chronic conditions. Methods Inf Med 2017;56(5):391-400 [FREE Full text] [doi: 10.3414/ME16-01-0135] [Medline: 29582934]
4. Lappenschaar M, Hommersom A, Lucas PJ, Lagro J, Visscher S, Korevaar JC, et al. Multilevel temporal Bayesian networks can model longitudinal change in multimorbidity. J Clin Epidemiol 2013 Dec;66(12):1405-1416 [FREE Full text] [doi: 10.1016/j.jclinepi.2013.06.018] [Medline: 24035172]
5. Pugh MJ, Swan AA, Carlson KF, Jaramillo CA, Eapen BC, Dillahunt-Aspillaga C, Trajectories of Resilience and Complex Comorbidity Study Team. Traumatic brain injury severity, comorbidity, social support, family functioning, and community reintegration among veterans of the Afghanistan and Iraq wars. Arch Phys Med Rehabil 2018 Feb;99(2S):S40-S49. [doi: 10.1016/j.apmr.2017.05.021] [Medline: 28648681]
6. Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A 2007 May 22;104(21):8685-8690 [FREE Full text] [doi: 10.1073/pnas.0701361104] [Medline: 17502601]
7. Loscalzo J, Kohane I, Barabasi A. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol 2007;3:124 [FREE Full text] [doi: 10.1038/msb4100163] [Medline: 17625512]
8. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. Circ Res 2012 Jul 20;111(3):359-374 [FREE Full text] [doi: 10.1161/CIRCRESAHA.111.258541] [Medline: 22821909]
9. Faruqui SH, Alaeddini A, Jaramillo CA, Potter JS, Pugh MJ. Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal Bayesian network. PLoS One 2018;13(7):e0199768 [FREE Full text] [doi: 10.1371/journal.pone.0199768] [Medline: 30001371]
10. Pearl J. Probabilistic Reasoning In Intelligent Systems: Networks Of Plausible Inference. San Francisco, California, USA: Morgan Kaufmann; 2014.
11. Heckerman D. A tutorial on learning with Bayesian networks. In: Holmes DE, editor. Innovations in Bayesian Networks: Theory and Applications. New York, USA: Springer; 2008:33-82.
12. Gore R, Reynolds PF. Applying Causal Inference to Understand Emergent Behavior. In: Proceedings of the Winter Simulation Conference. 2008 Presented at: WSC'08; December 7-10, 2008; Miami, FL, USA p. 712. [doi: 10.1109/wsc.2008.4736133]
13. Cai Z, Si S, Chen C, Zhao Y, Ma Y, Wang L, et al. Analysis of prognostic factors for survival after hepatectomy for hepatocellular carcinoma based on a Bayesian network. PLoS One 2015;10(3):e0120805 [FREE Full text] [doi: 10.1371/journal.pone.0120805] [Medline: 25826337]
14. Maglogiannis I, Zafiropoulos E, Platis A, Lambrinoudakis C. Risk analysis of a patient monitoring system using Bayesian network modeling. J Biomed Inform 2006 Dec;39(6):637-647 [FREE Full text] [doi: 10.1016/j.jbi.2005.10.003] [Medline: 16337837]
15. Gatti E, Luciani D, Stella F. A continuous time Bayesian network model for cardiogenic heart failure. Flex Serv Manuf J 2011 Dec 8;24(4):496-515. [doi: 10.1007/s10696-011-9131-2]
16. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual J, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 2006 May 19;125(4):801-814 [FREE Full text] [doi: 10.1016/j.cell.2006.03.032] [Medline: 16713569]
17. Lu X, Jain VV, Finn PW, Perkins DL. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. Mol Syst Biol 2007;3:98 [FREE Full text] [doi: 10.1038/msb4100138] [Medline: 17437023]
18. Sandri M, Berchialla P, Baldi I, Gregori D, de Blasi RA. Dynamic Bayesian networks to predict sequences of organ failures in patients admitted to ICU. J Biomed Inform 2014 Apr;48:106-113 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.008] [Medline: 24361388]
19. Liu Y, Safavi T, Dighe A, Koutra D. Graph summarization methods and applications: a survey. ACM Comput Surv 2018 Jul 16;51(3):1-34. [doi: 10.1145/3186727]
20. Jianbo S, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Machine Intell 2000;22(8):888-905. [doi: 10.1109/34.868688]
21. Ng AY, Jordan MI, 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, British Columbia, Canada p. 849-856 URL: https://dl.acm.org/doi/10.5555/2980539.2980649 [doi: 10.5555/2980539.2980649]
22. Toivonen H, Zhou F, Hartikainen A, Hinkka A. Compression of Weighted Graphs. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011 Presented at: KDD'11; August 21-24, 2011; San Diego, California p. 965-973. [doi: 10.1145/2020408.2020566]
23. Dhillon IS, Mallela S, Modha DS. Information-Theoretic Co-Clustering. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2003 Presented at: KDD'03; August 24-27, 2003; Washington, DC p. 89-98. [doi: 10.1145/956750.956764]

24. Chakrabarti D, Papadimitriou S, Modha D, Faloutsos C. Fully Automatic Cross-Associations. 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, Washington p. 79-88. [doi: 10.1145/1014052.1014064]
25. Chierichetti F, Kumar R, Lattanzi S, Mitzenmacher M, Panconesi A, Raghavan P. On Compressing Social Networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009 Presented at: KDD'09; June 28-July 1, 2009; Paris, France p. 219-228. [doi: 10.1145/1557019.1557049]
26. Lim Y, Kang U, Faloutsos C. SlashBurn: graph compression and mining beyond caveman communities. IEEE Trans Knowl Data Eng 2014 Dec 1;26(12):3077-3089. [doi: 10.1109/tkde.2014.2320716]
27. le Fevre K, Terzi E. GraSS: Graph Structure Summarization. In: Proceedings of the SIAM International Conference on Data Mining. 2010 Presented at: SDM'10; April 29-May 1, 2010; Columbus, Ohio, USA. [doi: 10.1137/1.9781611972801.40]
28. Purohit M, Prakash B, Kang C, Zhang Y, Subrahmanian V. Fast Influence-Based Coarsening for Large Networks. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014 Presented at: KDD'14; August 24-27, 2014; New York, USA p. 1296-1305. [doi: 10.1145/2623330.2623701]
29. Low Y, Bickson D, Gonzalez J, Guestrin C, Kyrola A, Hellerstein J. Distributed GraphLab: a framework for machine learning in the cloud. Proc VLDB Endow 2012 Apr 26;5(8):716-727 [FREE Full text] [doi: 10.14778/2212351.2212354]
30. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E 2004 Feb 26;69(2):26113. [doi: 10.1103/physreve.69.026113]
31. Yang J, Leskovec J. Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. 2013 Presented at: WSDM'13; February 4-8, 2013; Rome, Italy p. 587-596. [doi: 10.1145/2433396.2433471]
32. Karypis G, Kumar V. Multilevelk-way partitioning scheme for irregular graphs. J Parallel Distr Com 1998 Jan;48(1):96-129. [doi: 10.1006/jpdc.1997.1404]
33. Zeqian S, Kwan-Liu M, Eliassi-Rad T. Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE Trans Visual Comput Graphics 2006 Nov;12(6):1427-1439. [doi: 10.1109/tvcg.2006.107] [Medline: 17073366]
34. Li CT, Lin SD. Egocentric Information Abstraction for Heterogeneous Social Networks. In: Proceedings of the International Conference on Advances in Social Network Analysis and Mining. 2009 Presented at: ASONAM'09; July 20-22, 2009; Athens, Greece. [doi: 10.1109/asonam.2009.38]
35. Koutra D, Kang U, Vreeken J, Faloutsos C. Summarizing and understanding large graphs. Stat Anal Data Min 2015 May 18;8(3):183-202. [doi: 10.1002/sam.11267]
36. Miettinen P, Vreeken J. MDL4BMF: minimum description length for boolean matrix factorization. ACM Trans Knowl Discov Data 2014 Oct 7;8(4):1-31. [doi: 10.1145/2601437]
37. Miettinen P, Vreeken J. Model Order Selection for Boolean Matrix Factorization. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011 Presented at: KDD'11; August 21-24, 2011; San Diego, California p. 51-59. [doi: 10.1145/2020408.2020424]
38. Maccioni A, Abadi D. Scalable Pattern Matching over Compressed Graphs via Dedensification. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, USA p. 1755-1764. [doi: 10.1145/2939672.2939856]
39. Buehrer G, Chellapilla K. A Scalable Pattern Mining Approach to Web Graph Compression With Communities. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. 2008 Presented at: WSDM'08; February 11-12, 2008; California, USA p. 95-106. [doi: 10.1145/1341531.1341547]
40. Cook D, Holder L. Substructure discovery using minimum description length and background knowledge. J Artif Intell Res 1994 Feb 1;1:231-255 [FREE Full text] [doi: 10.1613/jair.43]
41. Akoglu L, McGlohon M, Faloutsos C. Oddball: spotting anomalies in weighted graphs. In: Pal N, editor. Advanced Techniques in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer; 2010:410-421.
42. Kang U, Tsourakakis C, Faloutsos C. PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations. In: Proceedings of the International Conference on Data Mining. 2009 Presented at: ICDM'09; December 6-9, 2009; Miami, FL, USA p. 229-238. [doi: 10.1109/icdm.2009.14]
43. Chan H, Akoglu L, Tong H. Make It or Break It: Manipulating Robustness in Large Networks. In: Proceedings of the 2014 SIAM International Conference on Data Mining. 2014 Presented at: SIAM'14; April 24-26, 2014; Pennsylvania, USA p. 325-333. [doi: 10.1137/1.9781611973440.37]
44. Tong H, Prakash B, Eliassi-Rad T, Faloutsos M, Faloutsos C. Gelling, and Melting, Large Graphs by Edge Manipulation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012 Presented at: CIKM'12; October 29-November 2, 2012; Hawaii, USA. [doi: 10.1145/2396761.2396795]
45. Chung F, Graham F. Spectral Graph Theory. Rhode Island, New York: American Mathematical Society; 1997.
46. Harary F, Schwenk A. The spectral approach to determining the number of walks in a graph. Pacific J Math 1979;80(2):443-449. [doi: 10.2140/pjm.1979.80.443]
47. Prakash BA, Chakrabarti D, Valler NC, Faloutsos M, Faloutsos C. Threshold conditions for arbitrary cascade models on arbitrary networks. Knowl Inf Syst 2012 Jul 7;33(3):549-575. [doi: 10.1007/s10115-012-0520-y]

48. Yang W, Chakrabarti D, Chenxi W, Faloutsos C. Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint. In: Proceedings of the 22nd International Symposium on Reliable Distributed Systems. 2003 Presented at: RELDIS'03; October 6-8, 2003; Florence, Italy. [doi: 10.1109/reldis.2003.1238052]
49. Cooper GF. A diagnostic method that uses causal knowledge and linear programming in the application of Bayes' formula. Comput Methods Programs Biomed 1986 Apr;22(2):223-237. [doi: 10.1016/0169-2607(86)90024-6] [Medline: 3519071]
50. Pearl J. Causality: Models, Reasoning and Inference. Second Edition. New York, USA: Cambridge University Press; 2009.
51. Darwiche A. Modeling and Reasoning with Bayesian Networks. New York, USA: Cambridge University Press; 2009.
52. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992 Oct;9(4):309-347. [doi: 10.1007/BF00994110]
53. Alon N. Eigenvalues and expanders. Combinatorica 1986 Jun 1;6(2):83-96 [FREE Full text] [doi: 10.1007/bf02579166]
54. Aldous D, Fill JA. Statistics at UC Berkeley: Department of Statistics. 2002. Reversible Markov Chains and Random Walks on Graphs URL: https://www.stat.berkeley.edu/~aldous/RWG/book.pdf [accessed 2020-04-09]
55. Chaiken S, Kleitman DJ. Matrix tree theorems. J Comb Theory A 1978 May;24(3):377-381. [doi: 10.1016/0097-3165(78)90067-5]
56. Chung F. Laplacians and the cheeger inequality for directed graphs. Ann Comb 2005 Apr;9(1):1-19. [doi: 10.1007/s00026-005-0237-z]
57. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006 Mar 28;65(1):31-78. [doi: 10.1007/s10994-006-6889-7]
58. Kahn AB. Topological sorting of large networks. Commun ACM 1962;5(11):558-562. [doi: 10.1145/368996.369025]
59. Tarjan R. Depth-first search and linear graph algorithms. SIAM J Comput 1972 Jun;1(2):146-160. [doi: 10.1137/0201010]
60. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, London: Oxford University Press; 2003.
61. Li L, Tong H, Xiao Y, Fan W. Cheetah: Fast Graph Kernel Tracking on Dynamic Graphs. In: Proceedings of the 2015 SIAM International Conference on Data Mining. 2015 Presented at: SIAM'15; April 30-May 2, 2015; Vancouver, BC, Canada p. 280-288. [doi: 10.1137/1.9781611974010.32]
62. Gunn JM, Ayton DR, Densley K, Pallant JF, Chondros P, Herrman HE, et al. The association between chronic illness, multimorbidity and depressive symptoms in an Australian primary care cohort. Soc Psychiatry Psychiatr Epidemiol 2012 Feb;47(2):175-184. [doi: 10.1007/s00127-010-0330-z] [Medline: 21184214]
63. Bay E, Hagerty BM, Williams RA, Kirsch N, Gillespie B. Chronic stress, sense of belonging, and depression among survivors of traumatic brain injury. J Nurs Scholarsh 2002;34(3):221-226. [doi: 10.1111/j.1547-5069.2002.00221.x] [Medline: 12237983]
64. Bhattacharya R, Shen C, Wachholtz AB, Dwibedi N, Sambamoorthi U. Depression treatment decreases healthcare expenditures among working age patients with comorbid conditions and type 2 diabetes mellitus along with newly-diagnosed depression. BMC Psychiatry 2016 Jul 19;16:247 [FREE Full text] [doi: 10.1186/s12888-016-0964-9] [Medline: 27431801]
65. Bramoweth AD, Renqvist JG, Hanusa BH, Walker JD, Germain A, Atwood CW. Identifying the demographic and mental health factors that influence insomnia treatment recommendations within a veteran population. Behav Sleep Med 2017 May 2 epub ahead of print. [doi: 10.1080/15402002.2017.1318752] [Medline: 28463021]
66. Concato J, Shah N, Horwitz R. Randomized, controlled trials, observational studies, and the hierarchy of research designs. In: Elliott D, Stern JE, editors. Research Ethics. New Hampshire, United States: Institute for the Study of Applied and Professional Ethics; 2017:207-212.
67. Corson K, Denneson LM, Bair MJ, Helmer DA, Goulet JL, Dobscha SK. Prevalence and correlates of suicidal ideation among operation enduring freedom and operation Iraqi freedom veterans. J Affect Disord 2013 Jul;149(1-3):291-298. [doi: 10.1016/j.jad.2013.01.043] [Medline: 23531358]
68. Diederichs C, Berger K, Bartels DB. The measurement of multiple chronic diseases-a systematic review on existing multimorbidity indices. J Gerontol A Biol Sci Med Sci 2011 Mar;66(3):301-311. [doi: 10.1093/gerona/glq208] [Medline: 21112963]
69. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. J Chronic Dis 1970 Dec;23(7):455-468. [doi: 10.1016/0021-9681(70)90054-8] [Medline: 26309916]
70. Fenton B, Goulet J, Bair M, Cowley T, Kerns R. Relationships between temporomandibular disorders, MSD conditions, and mental health comorbidities: findings from the veterans musculoskeletal disorders cohort. Pain Med 2018 Sep 1;19(Suppl 1):S61-S68. [doi: 10.1093/pm/pny145] [Medline: 30203016]
71. Glenn MB, O'Neil-Pirozzi T, Goldstein R, Burke D, Jacob L. Depression amongst outpatients with traumatic brain injury. Brain Inj 2001 Sep;15(9):811-818. [doi: 10.1080/02699050010025777] [Medline: 11516349]
72. Haibach JP, Haibach MA, Hall KS, Masheb RM, Little MA, Shepardson RL, et al. Military and veteran health behavior research and practice: challenges and opportunities. J Behav Med 2017 Feb;40(1):175-193. [doi: 10.1007/s10865-016-9794-y] [Medline: 27678001]
73. Hunter G, Yoon J, Blonigen DM, Asch SM, Zulman DM. Health care utilization patterns among high-cost VA patients with mental health conditions. Psychiatr Serv 2015 Sep;66(9):952-958. [doi: 10.1176/appi.ps.201400286] [Medline: 25930040]

74. Magnavita N, Garbarino S. Sleep, health and wellness at work: a scoping review. Int J Environ Res Public Health 2017 Nov 6;14(11):1347. [doi: 10.3390/ijerph14111347] [Medline: 29113118]
75. Mastrocola EL, Taylor AK, Chew-Graham C. Access to healthcare for long-term conditions in women involved in street-based prostitution: a qualitative study. BMC Fam Pract 2015 Sep 3;16:118 [FREE Full text] [doi: 10.1186/s12875-015-0331-9] [Medline: 26338724]
76. McGlinchey RE, Milberg WP, Fonda JR, Fortier CB. A methodology for assessing deployment trauma and its consequences in OEF/OIF/OND veterans: the TRACTS longitudinal prospective cohort study. Int J Methods Psychiatr Res 2017 Sep;26(3):e1556 [FREE Full text] [doi: 10.1002/mpr.1556] [Medline: 28211592]
77. Peterson J, Brommelsiek M, Amelung SK. An interprofessional education project to address veterans' healthcare needs. Int J High Educ 2016 Nov 3;6(1):1. [doi: 10.5430/ijhe.v6n1p1]
78. Rosenthal M, Christensen BK, Ross TP. Depression following traumatic brain injury. Arch Phys Med Rehabil 1998 Jan;79(1):90-103. [doi: 10.1016/s0003-9993(98)90215-5]
79. Swartz JA. Chronic medical conditions among jail detainees in residential psychiatric treatment: a latent class analysis. J Urban Health 2011 Aug;88(4):700-717 [FREE Full text] [doi: 10.1007/s11524-011-9554-9] [Medline: 21394659]
80. Thabrew H, Stasiak K, Hetrick S, Donkin L, Huss J, Highlander A, et al. Psychological therapies for anxiety and depression in children and adolescents with long-term physical conditions. Cochrane Database Syst Rev 2018 Dec 22;12:CD012488 [FREE Full text] [doi: 10.1002/14651858.CD012488.pub2] [Medline: 30578633]
81. Zis P, Daskalaki A, Bountouni I, Sykioti P, Varrassi G, Paladini A. Depression and chronic pain in the elderly: links and management challenges. Clin Interv Aging 2017;12:709-720 [FREE Full text] [doi: 10.2147/CIA.S113576] [Medline: 28461745]
82. Murray BS, Choe SE, Woods M, Ryan TE, Liu W. An in silico analysis of microRNAs: mining the miRNAome. Mol Biosyst 2010 Oct;6(10):1853-1862. [doi: 10.1039/c003961f] [Medline: 20539892]
83. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001 May 3;411(6833):41-42. [doi: 10.1038/35075138] [Medline: 11333967]
84. Pradhan M, Provan G, Middleton B, Henrion M. Knowledge Engineering for Large Belief Networks. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence. 1994 Presented at: UAI'94; July 29-31, 2994; Seattle, WA p. 484-490. [doi: 10.1016/b978-1-55860-332-5.50066-3]
85. Chen C, Tong H. Fast Eigen-Functions Tracking on Dynamic Graphs. In: Proceedings of the International Conference on Data Mining. 2015 Presented at: SIAM'15; April 30-May 2, 2015; Vancouver, BC, Canada. [doi: 10.1137/1.9781611974010.63]

Abbreviations

AIC: Akaike information criterion
AUC: area under the curve
BaPa: back pain
BN: Bayesian network
DAG: directed acyclic graph
Depr: depression
DFS: depth-first search
EAGL: eigenvalue analysis of the graph Laplacian
MCC: multiple chronic conditions
MWST: maximum weight spanning tree
PTSD: posttraumatic stress disorder
ROC: receiver operating characteristic
SuAb: substance abuse
TBI: traumatic brain injury
VA: Veterans Health Administration

Edited by G Eysenbach; submitted 24.09.19; peer-reviewed by R Moghaddass, H Carretta, IC Jeong, P Giabbanelli, R Gore; comments to author 11.11.19; revised version received 06.01.20; accepted 22.03.20; published 17.06.20.

Please cite as:
Faruqui SHA, Alaeddini A, Chang MC, Shirinkam S, Jaramillo C, NajafiRad P, Wang J, Pugh MJ
Summarizing Complex Graphical Models of Multiple Chronic Conditions Using the Second Eigenvalue of Graph Laplacian: Algorithm Development and Validation
JMIR Med Inform 2020;8(6):e16372
URL: http://medinform.jmir.org/2020/6/e16372/
doi:10.2196/16372
PMID:32554376

Original Paper

Distributed Regression Analysis Application in Large Distributed Data Networks: Analysis of Precision and Operational Performance

Qoua Her[1], MSc, PharmD, MSPharmD; Jessica Malenfant[1], MPH; Zilu Zhang[1], MSc; Yury Vilk[1], PhD; Jessica Young[1], PhD; David Tabano[2,3], MA, PhD; Jack Hamilton[4], AB; Ron Johnson[5], MA; Marsha Raebel[2], PharmD; Denise Boudreau[5], PhD; Sengwee Toh[1], ScD

[1]Harvard Medical School, Harvard Pilgrim Health Care Institute, Boston, MA, United States
[2]Institute for Health Research, Kaiser Permanente Colorado, Denver, CO, United States
[3]Center for Observational Research and Data Science, Bristol-Meyers Squibb, Lawrenceville, NJ, United States
[4]Division of Research, Kaiser Permanete North California, Oakland, CA, United States
[5]Health Research Institute, Kaiser Permanente Washington, Seattle, WA, United States

Corresponding Author:
Qoua Her, MSc, PharmD, MSPharmD
Harvard Medical School
Harvard Pilgrim Health Care Institute
401 Park Drive, 4th Floor East
Boston, MA, 02215
United States
Phone: 1 617 867 4885
Email: qouaher@gmail.com

Abstract

Background: A distributed data network approach combined with distributed regression analysis (DRA) can reduce the risk of disclosing sensitive individual and institutional information in multicenter studies. However, software that facilitates large-scale and efficient implementation of DRA is limited.

Objective: This study aimed to assess the precision and operational performance of a DRA application comprising a SAS-based DRA package and a file transfer workflow developed within the open-source distributed networking software PopMedNet in a horizontally partitioned distributed data network.

Methods: We executed the SAS-based DRA package to perform distributed linear, logistic, and Cox proportional hazards regression analysis on a real-world test case with 3 data partners. We used PopMedNet to iteratively and automatically transfer highly summarized information between the data partners and the analysis center. We compared the DRA results with the results from standard SAS procedures executed on the pooled individual-level dataset to evaluate the precision of the SAS-based DRA package. We computed the execution time of each step in the workflow to evaluate the operational performance of the PopMedNet-driven file transfer workflow.

Results: All DRA results were precise ($<10^{-12}$), and DRA model fit curves were identical or similar to those obtained from the corresponding pooled individual-level data analyses. All regression models required less than 20 min for full end-to-end execution.

Conclusions: We integrated a SAS-based DRA package with PopMedNet and successfully tested the new capability within an active distributed data network. The study demonstrated the validity and feasibility of using DRA to enable more privacy-protecting analysis in multicenter studies.

(JMIR Med Inform 2020;8(6):e15073) doi:10.2196/15073

KEYWORDS

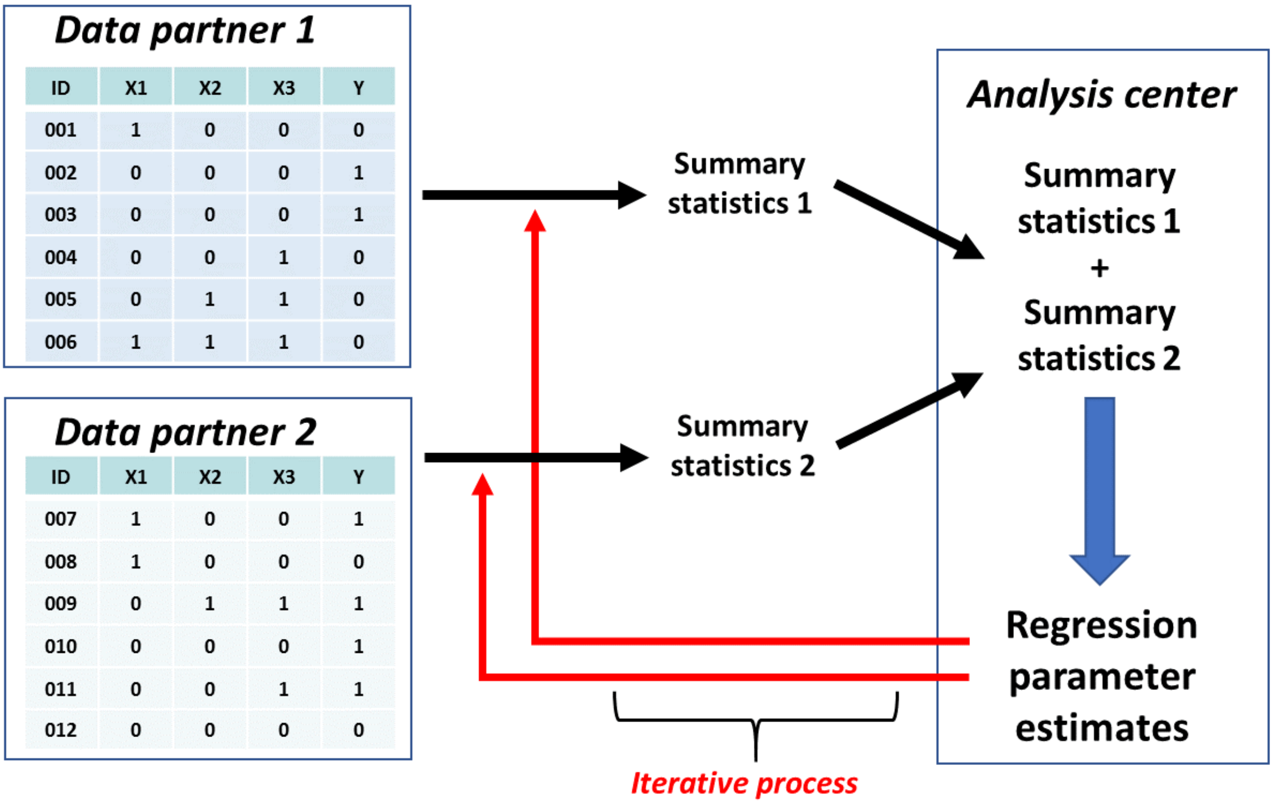
distributed regression analysis; distributed data networks; privacy-protecting analytics; pharmacoepidemiology; PopMedNet

Introduction

Background and Significance

Distributed regression analysis (DRA) is a suite of methods that perform multivariable regression analysis in multicenter studies without the need for pooling individual-level data [1,2]. Data partners compute highly summarized intermediate statistics (eg, sums of squares and cross products matrices) of their individual-level data and share these statistics with a trusted third-party or analysis center (Figure 1). The analysis center aggregates the intermediate statistics, assesses model convergence, and computes the regression parameter estimates. DRA is mathematically equivalent to the conventional regression analysis of pooled individual-level data. It achieves the same level of statistical sophistication using only summary-level information, thereby offering better protection for individual and institutional privacy without jeopardizing the scientific rigor of the analysis.

Figure 1. Distributed regression analysis with horizontally partitioned data.



Data partner 1

ID	X1	X2	X3	Y
001	1	0	0	0
002	0	0	0	1
003	0	0	0	1
004	0	0	1	0
005	0	1	1	0
006	1	1	1	0

Data partner 2

ID	X1	X2	X3	Y
007	1	0	0	1
008	1	0	0	0
009	0	1	1	1
010	0	0	0	1
011	0	0	1	1
012	0	0	0	0

However, DRA is not widely used in practice due to the operational challenges in implementing the approach [3]. The modeling process of common regression analyses (eg, logistic regression, Cox proportional hazards regression) is iterative and requires multiple exchanges of highly summarized intermediate statistics between the data partners and the analysis center. Manual execution of DRA is labor-intensive and highly susceptible to human errors (eg, transfer of incorrect files). There have been efforts to develop capabilities that coordinate and automate the iterative computation and file transfer process of DRA to make it a more practical analytical option in real-world multicenter studies [4-11]. These efforts have focused primarily on the programming language R and specially designed applications (eg, Java applets) to facilitate semiautomated or fully automated file transfers between the data partners and the analysis center [7-11]. The performance of these capabilities has largely been tested in simulated or relatively well-controlled environments [4-8], and no DRA application has been developed in SAS, another commonly used statistical software.

In our previous work, we enhanced PopMedNet, an open-source distributed networking software currently used by several large national distributed data networks (DDNs), to enable an automatable and iterative file transfer workflow for routine implementation of DRA [3]. This workflow coordinates and automates the iterative transfer of files between the data partners and the analysis center. We also created a SAS-based DRA package to conduct distributed linear, logistic, and Cox proportional hazards regression analysis in horizontally partitioned DDN [12,13], environments where each data partner holds information about distinct individuals [14,15]. We integrated the PopMedNet workflow with the SAS-based DRA package to create a DRA application.

Objectives

Despite the appealing theoretical properties of DRA, applications designed to perform the analysis can still be inoperable or produce biased results in real-world settings due to unappreciated factors (eg, human errors in execution, incompatible or different software versions, network or firewall restrictions, and network conditions). Evaluating the precision of DRA applications compared with the pooled individual-level data analysis and the feasibility of performing the analysis in reasonable execution times in real-world settings is needed to demonstrate DRA as a practical and valid analytical method. In this study, we demonstrate the feasibility of using the SAS-based DRA package and PopMedNet-driven file transfer workflow to perform DRA in a real-world horizontally partitioned DDN. Specifically, we quantify the precision of the SAS-based DRA package and the operational performance of the PopMedNet-driven file transfer workflow.

Methods

Study Setting: The Sentinel System

Funded by the US Food and Drug Administration, the Sentinel System is an active surveillance system designed to monitor the safety of approved medical products using longitudinal, regularly updated electronic health data from a network of 18 health plans and health care delivery systems [16,17]. Sentinel data partners transform their data into a common data model [18], which enables analytical programs and tools to be centrally developed and executed across data partners with minimal modifications. Over the years, the system has developed a suite of version-controlled, customizable, and freely available modular programs to rapidly query the transformed data across the DDN [19]. Among the tools is the Cohort Identification and Descriptive Analysis (CIDA) tool, a SAS program that assembles cohorts of individuals according to user-specified study parameters (eg, exposures, outcomes, inclusion and exclusion criteria) using established coding systems (eg, International Classification of Diseases, Ninth or Tenth Revision, Clinical Modification; National Drug Codes). The CIDA tool can generate a harmonized (ie, with the same covariates and covariate names) individual-level dataset at each data partner. Users can employ other tools (eg, Propensity Score Analysis Tool) or develop ad hoc analytical programs to query these datasets behind the data partner's firewall for complex inferential analyses.

Sentinel uses PopMedNet to facilitate file transfers between the data partners and the Sentinel Operations Center [20]. The Sentinel Operations Center, which serves as the analysis center for all Sentinel queries, uses a Web-based portal to create and securely distribute queries to data partners via PopMedNet. The data partners use a locally installed Microsoft Windows application, known as the DataMart Client, to retrieve the query and return the requested dataset, usually in aggregate-level format, to the Sentinel Operations Center. All file transfers between data partners and the Sentinel Operations Center are accomplished through secure HTTPS, secure sockets layer, or transport layer security connections. PopMedNet security and authentication requirements ensure that only approved queries are submitted to and responses returned by prespecified and approved data partners. In addition, the PopMedNet workflow is agnostic to query types, file formats (RData, sas, .docx, etc) and can transfer individual file sizes up to 2 GB.

SAS-Based Distributed Regression Analysis Application

There are numerous algorithms (eg, secure data integration, secure summation) for DRA in horizontally partitioned DDNs, environments where each data partner holds information about distinct patient cohorts [21,22]. In our previous work, we created a SAS-based DRA package comprising 2 interlinked SAS packages (one executed at the data partners and the other at the analysis center) using 2 algorithms: (1) distributed iteratively reweighted least squares to perform distributed linear and logistic regression analysis [12], and (2) distributed Newton-Raphson algorithm to perform distributed Cox proportional hazards regression analysis using the Efron or Breslow approximation for tied event times [13]. Both algorithms utilize a semitrusted third-party as the analysis center to aggregate the highly summarized intermediate statistics (eg, sums of squares and cross products matrices) and compute regression parameter estimates and SEs. We define a semitrusted third-party as a party that data partners trust with their summary-level information but not with their individual-level data. This party does not share data from any data partner with other data partners without consent, does not attempt to derive the individual-level data from the intermediate statistics, does not collude with data partners to derive any information about other data partners' individual-level data, and follows the DRA algorithms [23].

We provide a brief overview of the distributed iteratively reweighted least squares and the Newton-Raphson algorithms used to implement the SAS-based DRA package for distributed linear, logistic, and Cox proportional hazards regression analysis using the Sentinel Operations Center as the analysis center in Multimedia Appendix 1. A detailed description of these algorithms is available elsewhere [12,13].

PopMedNet Enhancements to Enable Automatable Distributed Regression Analysis

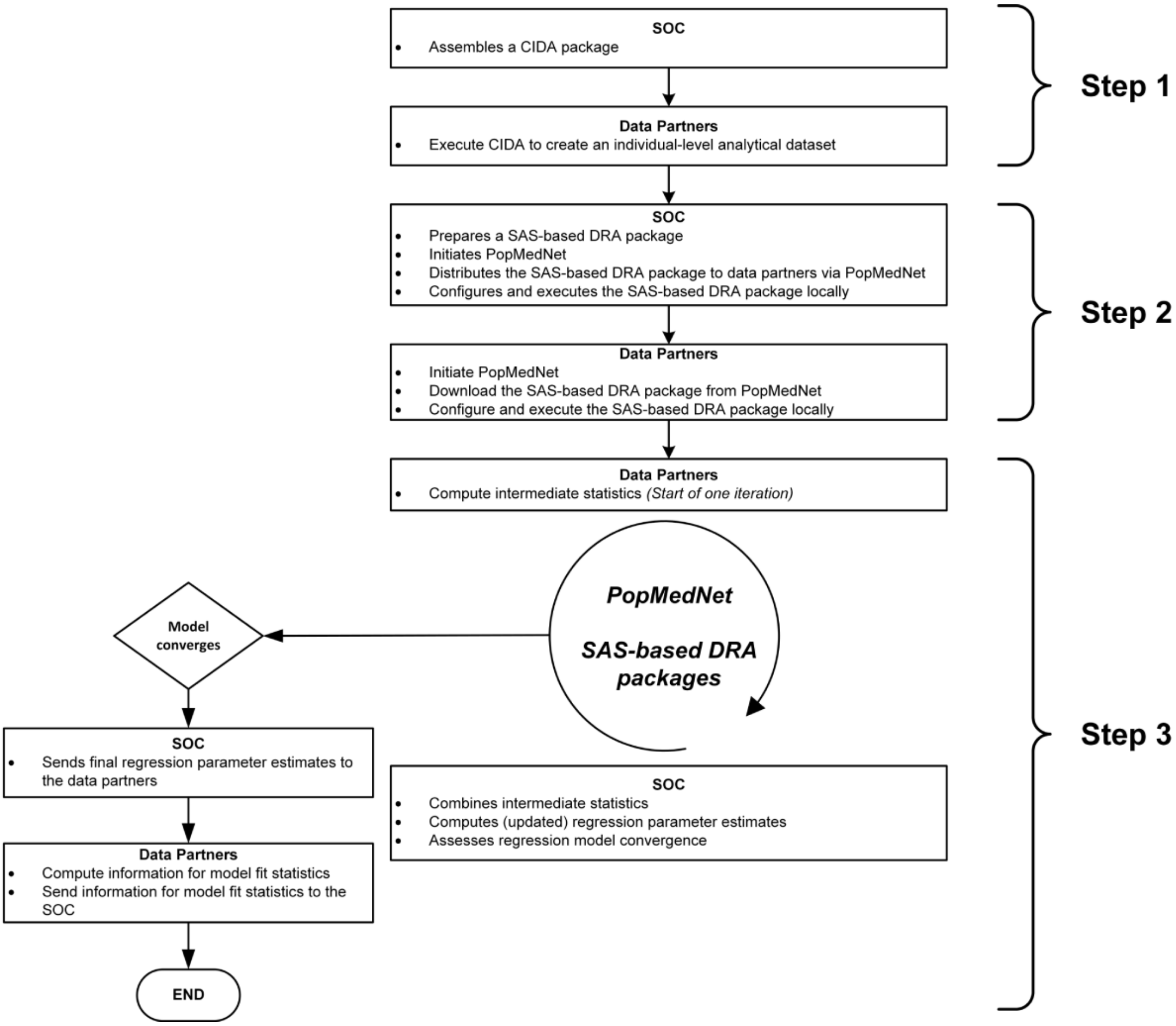
Both the distributed iteratively reweighted least squares and Newton-Raphson algorithms in the SAS-based DRA package utilize a master-worker process, where the analysis center directs the iterative DRA computations and the data partners execute these computations on their individual-level data with input (eg, updated regression parameter estimates) from the analysis center. Thus, an iterative file transfer workflow is required to transfer the highly summarized intermediate statistics and the updated regression parameter estimates between the data partners and the analysis center until the model converges or the analysis reaches a prespecified maximum number of iterations.

We previously enhanced PopMedNet to create an iterative and automatable file transfer workflow to facilitate routine DRA [3]. In brief, we built a back-end component, referred to as the *DRA-adapter*, into PopMedNet to allow the DataMart Client to upload files automatically and iteratively from and download files to prespecified folders at the data partners and the analysis center. We also developed functionalities for folder monitoring and trigger file creation and deletion in the DataMart Client to integrate the PopMedNet workflow with the two interlinked SAS packages of our SAS-based DRA package. A full description of the PopMedNet workflow and its integration with the SAS-based DRA package is available elsewhere [12,13]. We collectively refer to the integration of the SAS-based DRA package and the PopMedNet-driven file transfer workflow as the DRA application hereafter.

Distributed Regression Analysis: A 3-Step Process

A typical DRA includes 3 major steps [3]. Step 1 involves the assembly of a harmonized individual-level analytical dataset at each data partner. In step 2, the analysis center and each data partner execute a DRA algorithm locally. Step 3 involves the iterative transfer of the DRA algorithm outputs between the data partners and the analysis center until the regression model converges or the process reaches a prespecified maximum number of iterations. We used this 3-step process to guide our execution and evaluation of the DRA application with 3 Sentinel data partners, with the Sentinel Operations Center serving as the analysis center (Figure 2).

Figure 2. Three-step process to conduct distributed regression analysis with PopMedNet. CIDA: Cohort Identification and Descriptive Analysis Tool; DRA: Distributed Regression Analysis; SOC: Sentinel Operations Center.


SOC
Assembles a CIDA package
Data Partners
Execute CIDA to create an individual-level analytical dataset
Step 1
SOC
Prepares a SAS-based DRA package
Initiates PopMedNet
Distributes the SAS-based DRA package to data partners via PopMedNet
Configures and executes the SAS-based DRA package locally
Data Partners
Initiate PopMedNet
Download the SAS-based DRA package from PopMedNet
Configure and execute the SAS-based DRA package locally
Step 2
Data Partners
Compute intermediate statistics (Start of one iteration)
PopMedNet
SAS-based DRA packages
Model converges
SOC
Sends final regression parameter estimates to the data partners
SOC
Combines intermediate statistics
Computes (updated) regression parameter estimates
Assesses regression model convergence
Step 3
Data Partners
Compute information for model fit statistics
Send information for model fit statistics to the SOC
END

Step 1: Assemble a Harmonized Individual-Level Analytical Dataset at Each Data Partner

We used the CIDA tool (version 3.3.6) to assemble a harmonized individual-level analytical dataset of adult patients aged 18-79 years who received sleeve gastrectomy or Roux-en-Y gastric bypass in any care setting between January 1, 2010 and September 30, 2015 at 3 Sentinel data partners. To be eligible for cohort inclusion, patients must be continuously enrolled in a health plan with medical and drug coverage for at least 1 year before the index bariatric surgery, have at least one weight and height measurement that corresponded to a BMI ≥35 kg/m^2 in the year before surgery, and have at least one height and weight measurement in the year after surgery. We excluded patients with any bariatric procedure during the 1-year period before the index bariatric surgery. We also excluded patients with gastrointestinal cancer or a revised bariatric surgery procedure on the day of surgery. For each regression analysis, follow-up started on the day of the index bariatric surgery and continued until the occurrence of the outcome of interest (see below), death, end of health plan enrollment, or end of the study period. For distributed linear regression analysis, the outcome was a change in BMI within 1-year postsurgery, defined by subtracting the BMI measurement closest to the end of the

1-year postsurgery date from the last BMI measurement before surgery. For logistic regression, we created a binary outcome variable indicating *1* if the patient had weight loss ≥20% within 1-year postsurgery, and *0* if otherwise. For Cox regression analysis, we computed the time to weight loss ≥20% within the 1-year post-surgery period (Table 1).

Table 1. Analytical datasets and variables.

Regression model type	Outcome variable (within 1-year postsurgery)	Variables (exposure and confounders)
Linear	Change in BMI	Bariatric surgery exposure, age at surgery, sex, race and ethnicity, combined Charlson-Elixhauser comorbidity score, number of ambulatory visits, number of other ambulatory visits, number of inpatient stays, number of nonacute institutional stays, number of emergency department visits, BMI before bariatric surgery, number of days between last weight and height measurement and bariatric surgery, and data partner
Logistic	Weight loss ≥20%	Same as above
Cox	Time to weight loss ≥20%	Same as above

Step 2: Locally Execute the Distributed Regression Analysis Application at Each Data Partner and the Analysis Center

We assembled 3 separate SAS-based DRA packages to perform distributed linear, logistic, or Cox regression analyses and assessed the association between bariatric procedure (sleeve gastrectomy vs Roux-en-Y gastric bypass) and weight loss within 1-year postsurgery, adjusting for prespecified confounders (Table 1). For Cox regression analysis, we used the Efron approximation to handle tied event times. To be consistent with the standard SAS regression procedures, we prespecified a convergence criterion of <0.01 and a maximum of 25 iterations for distributed logistic and Cox regression analyses.

We distributed each SAS-based DRA package to the 3 data partners through PopMedNet (version 6.7). We instructed the data partners to (1) initiate the automated PopMedNet workflow, allowing the DataMart Client (version 6.7) to automatically download and unzip the SAS-based DRA package to a prespecified local directory, (2) manually place the individual-level analytical dataset created in step 1 in a prespecified local folder, (3) specify the file path to the SAS-based DRA package, and (4) execute the SAS-based DRA package in batch mode. Similarly, we instructed the Sentinel Operations Center to (1) initiate the automated PopMedNet workflow, (2) manually place the SAS-based DRA package for the analysis center in a prespecified local directory, (3) specify the file path to the SAS-based DRA package, and (4) execute the SAS-based DRA package in batch mode. Full details of these packages and examples of their execution have been previously described [12,13].

Step 3: Iteratively Transfer Distributed Regression Analysis Files Between the Data Partners and the Analysis Center

Once the data partners and the analysis center executed their SAS-based DRA package, the package ran continuously, awaiting input files (eg, updated regression parameter estimates or intermediate statistics) and DRA computation directions (eg, compute intermediate statistics, residuals, and SEs) from the Sentinel Operations Center. We used the PopMedNet workflow to transfer input files and computation directions iteratively and automatically between the data partners and the Sentinel Operations Center.

Evaluation of Precision and Operational Performance

We requested all data partners to securely transfer their deidentified individual-level analytical datasets to the Sentinel Operations Center. We assessed the precision of the SAS-based DRA package by comparing the DRA parameter estimates and SEs to those obtained from the pooled individual-level data analyses using standard SAS procedures. For distributed linear regression, we compared the model fit statistics R^2, Akaike information criterion (AIC), Sawa's Bayesian information criterion (BIC), and Schwarz BIC to the statistics obtained from a PROC REG run with the pooled individual-level data. For distributed logistic regression, we compared the model fit statistics log-likelihood, AIC, and Sawa's BIC to the statistics obtained from a PROC LOGISTIC run with the pooled individual-level data. For distributed Cox proportional hazards regression, we compared the model fit statistics log-likelihood, AIC, and Schwarz BIC to the statistics obtained from a PROC PHREG run with the pooled individual-level data. We considered the integration successful if the DRA parameter estimates and SEs and model fit statistics were precise to the results from the corresponding pooled individual-level data analyses (10^{-6}).

For distributed logistic regression, we also compared the receiver operating characteristic (ROC) curve and the area under the ROC curve with the corresponding curve and area obtained from a PROC LOGISTIC run with the pooled individual-level data. We considered the integration successful if the ROC curves were similar in likeliness and if the areas under the curves were comparable. To offer better privacy protection, we summarized individual-level predicted values for the distributed logistic regression analysis in bins of 6. Full details of this approximation method can be found elsewhere [12]. For distributed Cox proportional hazards analysis, we also compared the survival function curve with the curve obtained from a PROC PHREG run with the pooled individual-level data. We considered the integration successful if the survival function curves were similar in likeliness and if the median times to weight loss ≥20% were equivalent.

We extracted time stamps of status changes from PopMedNet and computed the average download, upload, SAS execution, and transfer time at the data partners and the analysis center to evaluate the operational performance of the DRA application. We also reported the average iteration time for each regression model type, and the time required to perform an end-to-end DRA in our test case.

We executed all SAS-based DRA packages in SAS versions 9.3 or 9.4, on a Windows desktop or server routinely used to perform Sentinel queries. All machines used to execute the SAS-based DRA packages and DataMart Client instance operated on a Windows 7 platform, with multiple Intel core processors ranging from 2.3 to 3.4 GHz, and 8 to 16 GB of RAM (Multimedia Appendix 2).

Results

Overview

We identified 5452 eligible patients among the 3 participating data partners (n_1=1706, n_2=2728, and n_3=1018). Of these, 981 patients received sleeve gastrectomy, whereas 4471 patients received Roux-en-Y gastric bypass during the study period. Within 1-year postsurgery, the BMI decreased on average by 9.8 kg/m^2 in sleeve gastrectomy patients and 18.7 kg/m^2 in Roux-en-Y gastric bypass patients. Five-hundred eighty-two of the 981 (59.3%) patients who had undergone sleeve gastrectomy and 3617 of the 4471 (80.10%) patients who had undergone Roux-en-Y gastric bypass had a weight loss ≥20% within the 1-year postsurgery period. The median time to a weight loss ≥20% was 223.9 days for patients who had undergone sleeve gastrectomy and 196.2 days for patients who had undergone Roux-en-Y gastric bypass.

Precision

Tables 2-4 summarize the precision of distributed linear, logistic, and Cox proportional hazards regression analyses. Table 5 shows the model fit statistics of the 3 regression models. All DRA parameter estimates, SEs, and model fit statistics were highly comparable to the estimates obtained from the pooled individual-level analyses that used standard SAS regression procedures. The ROC curve in distributed logistic regression (Figure 3) and the survival function in distributed Cox regression (Figure 4) were similar to those obtained from the pooled individual-level data analyses. The DRA application reported an area under the curve (AUC) of 0.6591 for logistic regression (vs 0.6592 from the pooled individual-level data analysis) and 184 days for Cox proportional hazards analysis (vs 184 days from the pooled individual-level data analysis) as the median time to weight loss ≥20%.

Table 2. Distributed linear regression vs pooled individual-level linear regression.

Covariates	Distributed regression analysis		Pooled individual-level analysis		Difference in parameter estimate	Difference in SE
	Parameter estimate	SE	Parameter estimate	SE		
Intercept	34.03935	0.61075	34.03935	0.61075	3.66 x 10^{-12}	−9.14 x 10^{-13}
Exposure	2.04714	0.28723	2.04714	0.28723	−4.15 x 10^{-13}	−4.30 x 10^{-13}
Age	−0.03334	0.00837	−0.03334	0.00837	−3.68 x 10^{-14}	−1.25 x 10^{-14}
Preindex BMI	−0.99983	0.00050	−0.99983	0.00050	−6.00 x 10^{-15}	−7.44 x 10^{-16}
Combined comorbidity score	0.04388	0.06949	0.04388	0.06949	3.59 x 10^{-15}	−1.04 x 10^{-13}
Number of ambulatory visits	−0.03068	0.01008	−0.03068	0.01008	−6.59 x 10^{-17}	−1.51 x 10^{-14}
Number of emergency department visits	0.10329	0.08749	0.10329	0.08749	−2.79 x 10^{-14}	−1.31 x 10^{-13}
Number of inpatient visits	0.88725	0.25976	0.88725	0.25976	−6.51 x 10^{-13}	−3.89 x 10^{-13}
Number of nonacute institutional stay	1.32338	1.79056	1.32338	1.79056	4.21 x 10^{-13}	−2.68 x 10^{-12}
Number of other ambulatory visits	0.02159	0.00873	0.02159	0.00873	1.22 x 10^{-14}	−1.31 x 10^{-14}
Days between BMI measurement and index procedure	0.01207	0.00567	0.01207	0.00567	3.92 x 10^{-15}	−8.48 x 10^{-15}
Race[a]						
Unknown	0.94212	0.26841	0.94212	0.26841	−4.16 x 10^{-13}	−4.02 x 10^{-13}
American Indian or Alaska Native	−0.30948	0.69817	−0.30948	0.69817	−2.39 x 10^{-13}	−1.04 x 10^{-12}
Asian	−0.16853	0.63001	−0.16853	0.63001	−4.52 x 10^{-13}	−9.42 x 10^{-13}
Black or African American	1.51961	0.29206	1.51961	0.29206	−9.95 x 10^{-14}	−4.37 x 10^{-13}
Native Hawaiian or other Pacific Islander	−1.22315	1.04973	−1.22315	1.04973	−4.11 x 10^{-13}	−1.57 x 10^{-12}
Female	−1.22366	0.23205	−1.22366	0.23205	−5.33 x 10^{-13}	−3.47 x 10^{-13}
Surgery year[a]						
2011	0.15150	0.30361	0.15150	0.30361	−5.94 x 10^{-13}	−4.54 x 10^{-13}
2012	−0.24904	0.30372	−0.24904	0.30372	−6.47 x 10^{-13}	−4.54 x 10^{-13}
2013	−0.02308	0.30223	−0.02308	0.30223	−6.08 x 10^{-13}	−4.52 x 10^{-13}
2014	0.32767	0.30609	0.32767	0.30609	−5.93 x 10^{-13}	−4.58 x 10^{-13}
2015	−0.25767	0.33352	−0.25767	0.33352	−6.18 x 10^{-13}	−4.99 x 10^{-13}
Data partner site[a]						
2	−1.10559	0.31373	−1.10559	0.31373	2.89 x 10^{-15}	−4.69 x 10^{-13}
3	−0.10990	0.30341	−0.10990	0.30341	−2.07 x 10^{-13}	−4.54 x 10^{-13}

[a] Reference groups: race (white), surgery year (2010), and data partner site (1).

Table 3. Distributed logistic regression vs pooled individual-level logistic regression.

Covariates	Distributed regression analysis		Pooled individual-level analysis		Difference in parameter estimate	Difference in SE
	Parameter estimate	SE	Parameter estimate	SE		
Intercept	2.11573	0.22833	2.11573	0.22833	−6.22 x 10^{-15}	−1.00 x 10^{-14}
Exposure	1.06711	0.09895	−1.06711	0.09895	−2.00 x 10^{-15}	−1.80 x 10^{-16}
Age	−0.01606	0.00316	−0.01607	0.00316	−4.51 x 10^{-17}	−1.57 x 10^{-16}
Preindex BMI	0.00003	0.00020	0.00003	0.00020	6.51 x 10^{-19}	2.44 x 10^{-19}
Combined comorbidity score	−0.02623	0.02561	−0.02623	0.02561	−6.97 x 10^{-16}	−3.12 x 10^{-17}
Number of ambulatory visits	0.01155	0.00447	0.01155	0.00447	6.25 x 10^{-17}	1.13 x 10^{-17}
Number of emergency department visits	−0.06230	0.03132	−0.06230	0.03133	3.05 x 10^{-16}	1.39 x 10^{-17}
Number of inpatient visits	−0.12098	0.08940	−0.12098	0.08940	1.75 x 10^{-15}	−2.36 x 10^{-16}
Number of nonacute institutional stay	0.42510	0.78809	0.42510	0.78809	−2.00 x 10^{-15}	−3.33 x 10^{-16}
Number of other ambulatory visits	0.00381	0.00340	0.00381	0.00340	3.17 x 10^{-17}	−2.91 x 10^{-17}
Days between BMI measurement and index procedure	−0.00266	0.00201	−0.00266	0.00201	3.90 x 10^{-17}	−4.77 x 10^{-18}
Race[a]						
Unknown	−0.39685	0.09485	−0.39685	0.09485	0.00 x 10^{+00}	−2.50 x 10^{-16}
American Indian or Alaska Native	−0.13938	0.26230	−0.13938	0.26230	−1.11 x 10^{-16}	5.55 x 10^{-17}
Asian	−0.37257	0.22341	−0.37257	0.22341	−3.04 x 10^{-14}	2.78 x 10^{-17}
Black or African American	−0.29617	0.10507	−0.29617	0.10507	−3.33 x 10^{-16}	−9.71 x 10^{-17}
Native Hawaiian or Other Pacific Islander	−0.02910	0.40543	−0.02910	0.40543	−6.14 x 10^{-16}	0.00 x 10^{+00}
Female	0.19993	0.08422	0.19993	0.08422	−1.80 x 10^{-15}	−3.61 x 10^{-16}
Surgery year[a]						
2011	−0.10269	0.11683	−0.10269	0.11684	6.37 x 10^{-15}	−5.55 x 10^{-17}
2012	0.05547	0.11897	0.05547	0.11897	5.45 x 10^{-15}	−1.67 x 10^{-16}
2013	−0.11956	0.11382	−0.11956	0.11382	6.80 x 10^{-15}	−1.94 x 10^{-16}
2014	−0.10956	0.11617	−0.10956	0.11617	4.36 x 10^{-15}	−1.80 x 10^{-16}
2015	0.03701	0.12798	0.03701	0.12798	6.47 x 10^{-15}	−2.50 x 10^{-16}
Data partner site[a]						
2	−0.10433	0.11751	−0.10433	0.11751	4.51 x 10^{-15}	−9.99 x 10^{-16}
3	0.75506	0.12577	0.75506	0.12577	2.11 x 10^{-15}	−2.50 x 10^{-16}

[a] Reference groups: Race (white), surgery year (2010), and data partner site (1).

Table 4. Distributed Cox proportional hazards regression vs pooled individual-level Cox proportional hazards regression.

Covariates	Distributed regression analysis		Pooled individual-level analysis		Difference in parameter estimate	Difference in SE
	Parameter estimate	SE	Parameter estimate	SE		
Exposure	−0.58160	0.05275	−0.58160	0.05275	6.66 x 10^{-16}	−8.33 x 10^{-17}
Age	−0.01107	0.00146	−0.01107	0.00146	1.39 x 10^{-17}	−9.11 x 10^{-18}
Preindex BMI	−0.00006	0.00009	−0.00006	0.00009	2.85 x 10^{-19}	−1.49 x 10^{-19}
Combined comorbidity score	−0.00787	0.01205	−0.00787	0.01205	−3.64 x 10^{-17}	−1.04 x 10^{-17}
Number of ambulatory visits	0.00584	0.00158	0.00584	0.00158	−2.95 x 10^{-17}	1.08 x 10^{-18}
Number of emergency department visits	−0.01873	0.01679	−0.01873	0.00158	1.56 x 10^{-16}	−2.43 x 10^{-17}
Number of inpatient visits	−0.08587	0.04580	−0.08587	0.04580	−9.58 x 10^{-16}	−1.25 x 10^{-16}
Number of nonacute institutional stay	0.06626	0.29266	0.06626	0.29266	3.75 x 10^{-16}	−3.33 x 10^{-16}
Number of other ambulatory visits	0.00279	0.00134	0.00279	0.00134	4.03 x 10^{-17}	−1.52 x 10^{-18}
Days between BMI measurement and index procedure	−0.00221	0.00096	−0.00221	0.00096	2.39 x 10^{-17}	−2.17 x 10^{-18}
Race[a]						
Unknown	−0.18898	0.04765	−0.18898	0.04765	5.27 x 10^{-16}	0.00 x 10^{+00}
American Indian or Alaska Native	−0.07476	0.12019	−0.07476	0.12019	1.25 x 10^{-16}	2.78 x 10^{-17}
Asian	−0.22309	0.10933	−0.22309	0.10933	−2.78 x 10^{-17}	6.94 x 10^{-17}
Black or African American	−0.18457	0.05116	−0.18457	0.05116	1.94 x 10^{-16}	−1.39 x 10^{-17}
Native Hawaiian or Other Pacific Islander	−0.19748	0.17333	−0.19748	0.17333	1.42 x 10^{-15}	2.78 x 10^{-17}
Female	−0.00887	0.04052	−0.00887	0.04052	−1.24 x 10^{-15}	−3.47 x 10^{-17}
Surgery year[a]						
2011	−0.08021	0.05176	−0.08021	0.05176	8.60 x 10^{-16}	1.11 x 10^{-16}
2012	−0.02547	0.05136	−0.02547	0.05136	4.61 x 10^{-16}	7.63 x 10^{-17}
2013	−0.09519	0.05195	−0.09519	0.05195	1.17 x 10^{-15}	4.86 x 10^{-17}
2014	−0.16866	0.05235	−0.16866	0.05235	8.60 x 10^{-16}	1.18 x 10^{-16}
2015	0.24763	0.05640	0.24763	0.05640	3.89 x 10^{-16}	1.04 x 10^{-16}
Data partner site[a]						
2	−0.15270	0.05188	−0.15270	0.05188	2.11 x 10^{-15}	-6.94 x 10^{-18}
3	0.33440	0.05161	0.33440	0.05161	8.33 x 10^{-16}	2.08 x 10^{-17}

[a] Reference groups: race (white), surgery year (2010), and data partner site (1).

Table 5. Comparison of model fit statistics between distributed regression and pooled individual-level data analysis.

Regression model type and statistic or test	Distributed regression analysis	Pooled individual-level data analysis	Difference in model fit statistics
Linear			
R^2	0.9987	0.9987	3.89×10^{-15}
Akaike information criterion	20089.6538	20089.6538	-1.59×10^{-08}
Sawa's Bayesian information criterion	20091.8710	20091.8710	-1.59×10^{-08}
Schwarz's Bayesian information criterion	20247.5868	20247.5868	-1.59×10^{-08}
Logistic			
-2 log-likelihood	5423.2491	5423.2491	1.36×10^{-11}
Akaike information criterion	5471.2491	5471.2491	1.36×10^{-11}
Sawa's Bayesian information criterion	5629.5265	5629.5265	1.36×10^{-11}
Area under the ROC[a] curve	0.6591	0.6592	-1.00×10^{-04}
Hosmer-Lemeshow (chi-square statistics)	1.3405	1.5596	-2.19×10^{-01}
Hosmer-Lemeshow, *P* value (*df*)	.995 (8)	.991 (8)	3.38×10^{-03}
Cox			
-2 log-likelihood	66217.7270	66217.7270	1.46×10^{-11}
Akaike information criterion	66263.7270	66263.7270	1.46×10^{-11}
Schwarz's Bayesian information criterion	66409.6070	66409.6070	1.46×10^{-11}
Median time to event (days)	184	184	0

[a] ROC: receiver operating characteristic.

Figure 3. Comparison of receiver operating characteristic curves between distributed logistic regression (left) and pooled individual-level logistic regression (right). To offer better privacy-protecting, individual-level predicted values were summarized in bins of 6 and transferred to the analysis center for aggregation in the distributed logistic regression analysis. The size of the bin is user-specified. ROC: receiver operating characteristic.

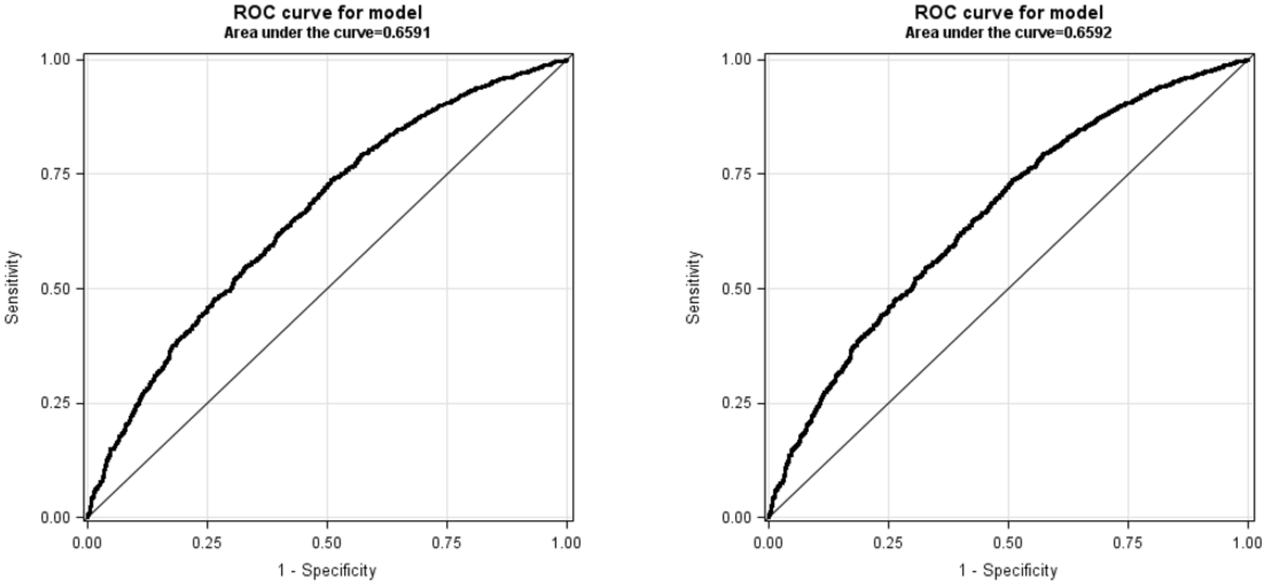
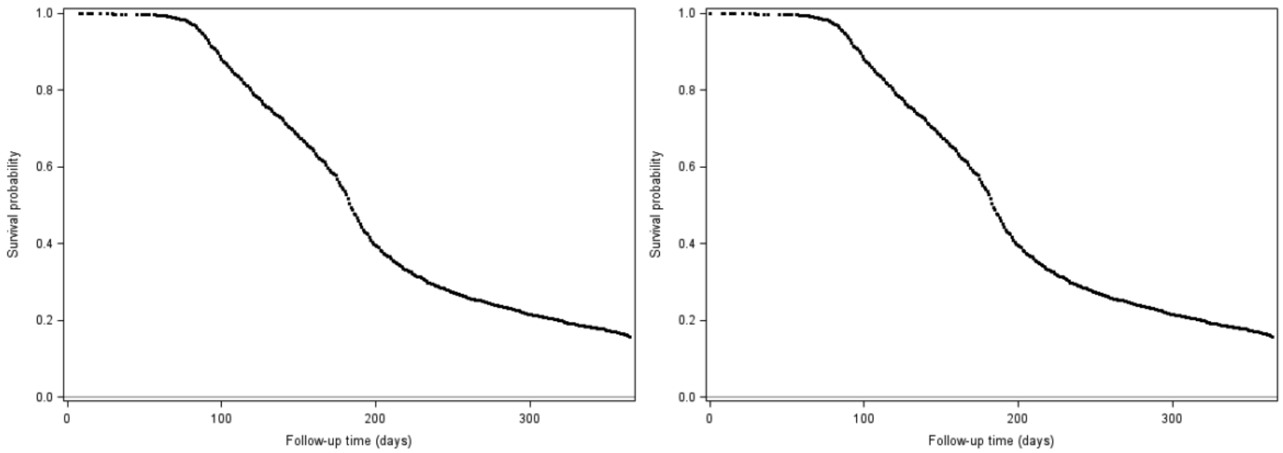


Figure 4. Comparison of survival functions between distributed cox proportional hazards regression (left) and pooled individual-level cox proportional hazards regression (right). The survival curves were evaluated at the mean value of covariates for patients with events.



Operational Performance

As expected, the closed-form solution of distributed linear regression analysis required only two iterations, one for computing the regression parameter estimates and SEs and the other for computing the model fit statistics. Both logistic and Cox proportional hazards regression analyses required 6 iterations for model convergence in our test case. Each file transfer process transferred between 3 and 10 files with sizes of 1 to 800 KB.

We extracted 111, 271, and 271 time stamps of status changes from PopMedNet for distributed linear, logistic, and Cox analysis, respectively. Table 6 summarizes the operational performance of the DRA application. It took an average of 102.4 s to complete one DRA iteration across all regression model types. The file transfer workflow (file upload, download, and transfer to the reciprocal party) accounted for 89% of the iteration time. Downloading and uploading the DRA files at the Sentinel Operations Center required an average of 28.6 and 9.8 s, respectively. File transfer from the Sentinel Operations Center to the data partners took on average 9.4 s. Downloading and uploading the DRA files at the data partners required an average of 10.1 and 15.5 s, respectively. File transfer from the data partners to the Sentinel Operations Center took an average 22.1 s. Computing the intermediate statistics at the data partners required an average of 8.0 s, whereas computing the updated regression parameters took an average of 3.8 s at the Sentinel Operations Center.

The distributed Cox regression required the greatest amount of iteration time (113.5 s), followed by logistic regression (95.0 s) and linear regression (91.5 s). Overall, distributed linear regression analysis with our bariatric surgery test case required 440.7 s to complete, whereas logistic and Cox proportional hazards regression analysis required 925.5 and 1016.0 s, respectively.

Table 6. Operational performance of the distributed regression analysis application.

Performance metric	Linear	Logistic	Cox	Overall
Required number of iterations for model convergence	2	6	6	—[a]
Total run time	440.7	925.5	1,016.0	—
Average iteration time, mean (SE)	91.5 (10.5)	95 (3.1)	113.5 (5.2)	102.4 (3.8)
Sentinel operations center (analysis center)				
Average download time, mean (SE)	20.5 (5.4)	20.6 (1.3)	39.4 (4)	28.6 (3.2)
Average computation time, mean (SE)	4.3 (2.6)	3 (1.1)	4.4 (0.4)	3.8 (0.6)
Average upload time, mean (SE)	8.4 (1.1)	10.2 (0.7)	9.9 (0.6)	9.8 (0.4)
Average file transfer time (to data partners), mean (SE)	10.5 (0.4)	9.1 (0.5)	9.4 (0.5)	9.4 (0.3)
Data partners				
Average download time, mean (SE)	8.6 (1.2)	10.3 (0.6)	10.3 (0.8)	10.1 (0.4)
Average computation time, mean (SE)	8.2 (0.8)	7.9 (0.4)	8 (0.3)	8 (0.2)
Average upload time, mean (SE)	15.6 (1.2)	15.9 (0.6)	15.1 (0.3)	15.5 (0.3)
Average file transfer time (to analysis center), mean (SE)	20 (0.8)	21.8 (1.9)	23.1 (1.2)	22.1 (1.0)

[a] N/A: not applicable.

Discussion

Principal Findings

We have successfully integrated a SAS-based DRA package with PopMedNet, an open-source distributed networking software, and performed DRA in select data partners within a real-world DDN. Our application was able to compute regression parameters, SEs, model fit statistics, and model fit graphics of 3 regression model types (linear, logistic, and Cox proportional hazards) that were within machine precision or similar in likeliness to those produced using standard SAS regression procedures, without the need to share any individual-level data, in under 20 min. The study demonstrated the feasibility and validity of performing multivariable regression analysis in a multicenter setting while limiting the risk of disclosing sensitive individual or institutional information.

Previous Studies

Previous studies have used simulated or relatively well-controlled distributed environments to demonstrate the ability to perform DRA with only summary-level information [4-8]. These studies have consistently reported that DRA produced precise (generally $<10^{-12}$) results compared with the results from the pooled individual-level data analysis. However, information on the operational performance (computation and file transfer time) of DRA algorithms or workflows is scarce. The closest experience to our DRA application is a Web-based DRA software developed by the SCAlable National Network for Effectiveness Research (SCANNER) [11]. This software is composed of a network portal with a set of Web services and virtual machines that host data from data-contributing sites and several libraries of analytical programs. At the time of our analysis, 3 method libraries were available in the SCANNER software: a cohort discovery tool, an algorithm to perform meta-analyses with distributed data, and an algorithm to perform distributed logistic regression analysis (Grid Binary LOgistic Regression, GLORE) [6]. The authors reported that GLORE produced results equivalent to those from the pooled individual-level data analysis, and software response times of 0.015 s with a dataset of 580 records (with a binary outcome variable, a treatment indicator variable, and 24 covariates) and 27.02 s with a dataset of 10,000 records (with a binary outcome variable and 5 covariates) partitioned among 3 different institutions.

Our DRA application required significantly more time for model convergence than the SCANNER software. However, this additional time for model convergence may be considered marginal in practice, where other aspects of a multicenter study are typically more time-consuming. For example, developing a study protocol and analysis plan or assembling an analytical dataset at each participating data partner for DRA may require considerably more time than the time required to perform DRA. There are also several key differences between our application and the SCANNER software that may explain the difference in operational performance. Specifically, the SCANNER software requires users to install a virtual machine and open ports to the master node hosting the SCANNER hub. This design may have shorter file upload, transfer, and download times between the execution nodes, as files are only transferred between homogeneous virtual machines on the server and not subject to impediments such as firewall security protocols, additional workload, and upload, transfer, and download speeds.

The operational performance of the SCANNER software makes it a desirable option for DRA in networks that are amenable to installing the required software and applications. We previously found that most Sentinel data partners were unwilling to install new software or make modifications to their existing hardware configurations to perform DRA [3]. We chose to develop the DRA application using SAS and PopMedNet because all Sentinel data partners have both software in their systems. In addition, several other large DDNs, including the National Patient-Centered Clinical Research Network [24] and the National Institutes of Health's Health Care Systems Research Collaboratory [25], use PopMedNet as their file transfer software. In other words, our DRA application requires no new software installation or modifications to existing hardware configurations in DDNs that employ SAS as their statistical software and PopMedNet as their file transfer software. The 3 data partners that participated in this project are also members of numerous PopMedNet-based DDNs. Therefore, the successful integration of our SAS-based DRA package with PopMedNet and execution of DRA with these data partners have the potential to extend DRA beyond the Sentinel System.

Limitations

Our study is not without limitations. First, DRA requires infrastructure and processes beyond the algorithms and technology described in this paper. For example, DRA with our application requires harmonized individual-level datasets. Since its inception, Sentinel has continuously enhanced its common data model, routine analytical tools, and data quality assurance processes. Thus, Sentinel data partners can rapidly create harmonized analytical datasets for DRA. Research networks and investigators without the same infrastructure may not be able to perform DRA with our application as easily, even if data partners are willing to use PopMedNet as their data-sharing software.

Second, we tested the DRA application with only 3 Sentinel data partners, and all tests were completed in a Windows version of SAS (desktop or server). It is possible that different hardware configurations not found at these data partners or different versions of SAS (Linux or Unix) could change the precision and operational performance or even inhibit the execution of our DRA application. However, we previously found only 3 different configurations of the required hardware components (DataMart Client, SAS software, and the common folder structure) among Sentinel data partners [3]. All 3 hardware configurations were represented among the 3 data partners in this study. We also found the reconfiguration of these components to be relatively straightforward. Therefore, it may be possible to have data partners with other configurations make minor changes to implement our DRA application. During the development of the DRA application, we were able to successfully execute our application on a Linux server with a fourth data partner, by placing the application on a Linux server directory accessible to the DataMart Client as a mapped

Windows network drive. This allowed the DataMart Client to access the same file system as the DRA application. Overall, additional testing with more data partners with different hardware configurations and different versions of SAS is needed to fully ensure that our DRA application is operable across different DDNs, research networks, operation systems, and environments.

Third, our precision and operational performances were based on a small sample of successful end-to-end executions of our DRA application. These executions were limited to regression models with 23 variables and analytical datasets ranging from 1000 to 3000 patients distributed among 3 data partners. Future work should include more end-to-end executions, regression models with more variables, datasets of larger sample sizes, and more data partners. However, we found that 89% of the iteration time was attributed to file transfer time, which was largely driven by the number of files, size of the files transferred, and network conditions (upload, download, and transfer speeds, firewall security protocols, and workload). Because the files contain highly summarized information, increasing the number of variables or patients will not increase the number of files or substantially increase the size of the files to be transferred. In this study, each file transfer process transferred files that were less than 1 MB. Our internal testing of analyses with more variables, patients, and data partners did not result in file sizes larger than a few MBs or increased the iteration time. Thus, we do not anticipate DRA with more variables, patients, and data partners in a real-world setting to have a considerable impact on the operational performance of our DRA application. In addition, network conditions at each data partner can vary depending on the workload. We could not vary network conditions at each data partner to formally analyze its impact on the operational performance. However, we did complete our experiments with 3 Sentinel data partners, with machines that are routinely used to fulfill Sentinel query requests. Thus, our results on precision and operational performance likely represent what potential users of DRA will experience in practice.

Fourth, our bariatric surgery test case was relatively simplistic and not as sophisticated as an actual clinical or epidemiologic study. For example, we did not include all the potential confounders. Therefore, the results of our analysis did not have any causal interpretation.

Finally, although DRA uses the intermediate statistics at each data partner to perform multivariable regression analysis, the risk of reidentifying specific individuals is not 0. Under certain conditions (eg, uncommon individual attributes coded with indicator variables), there could be leakage of personal information that could be used to infer or identify specific individuals [26]. To further protect privacy, DRA can be performed using more secure algorithms, such as encrypting or perturbing the intermediate statistics. Future work should explore the integration of these more secure DRA algorithms into our DRA application.

Conclusions

We have successfully developed and integrated a SAS-based DRA package with an iterative and automatable PopMedNet-driven file transfer workflow to create a DRA application and conduct DRA in select data partners within a real-world DDN. The application produced results that were within machine precision to the results from the pooled individual-level data analyses using standard SAS regression procedures. The end-to-end execution times were reasonable, demonstrating that DRA can be a practical and valid analytical method in real-world settings.

Acknowledgments

This work was supported by the Office of the Assistant Secretary for Planning and Evaluation and the Food and Drug Administration (HHSF223201400030I/HHSF22301006T).

Conflicts of Interest

ST is the principal investigator of projects funded by the National Institutes of Health (U01EB023683) and the Agency for Healthcare Research and Quality (R01HS026214).

Multimedia Appendix 1
Distributed regression analysis algorithms.
[DOCX File , 42 KB - medinform_v8i6e15073_app1.docx]

Multimedia Appendix 2
Analysis center and data partner hardware description.
[DOCX File , 23 KB - medinform_v8i6e15073_app2.docx]

References

1. Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. J Comput Graph Stat 2005;14(2):263-279. [doi: 10.1198/106186005x47714]
2. Fienberg SE, Fulp WJ, Slavkovic AB, Wrobel TA. 'Secure' log-linear and logistic regression analysis of distributed databases. In: Domingo-Ferrer J, Franconi L, editors. Privacy in Statistical Databases. Berlin, Heidelberg: Springer; 2006:277-290.

3. Her QL, Malenfant JM, Malek S, Vilk Y, Young J, Li L, et al. A query workflow design to perform automatable distributed regression analysis in large distributed data networks. EGEMS (Wash DC) 2018 May 25;6(1):11 [FREE Full text] [doi: 10.5334/egems.209] [Medline: 30094283]
4. Jiang X, Wu Y, Marsolo K, Ohno-Machado L. Development of a web service for analysis in a distributed network. EGEMS (Wash DC) 2014;2(1):1053 [FREE Full text] [doi: 10.13063/2327-9214.1053] [Medline: 25848586]
5. Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 2010 Oct;39(5):1372-1382 [FREE Full text] [doi: 10.1093/ije/dyq111] [Medline: 20630989]
6. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;19(5):758-764 [FREE Full text] [doi: 10.1136/amiajnl-2012-000862] [Medline: 22511014]
7. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015 Nov;22(6):1212-1219 [FREE Full text] [doi: 10.1093/jamia/ocv083] [Medline: 26159465]
8. Jiang W, Li P, Wang S, Wu Y, Xue M, Ohno-Machado L, et al. WebGLORE: a web service for Grid LOgistic REgression. Bioinformatics 2013 Dec 15;29(24):3238-3240 [FREE Full text] [doi: 10.1093/bioinformatics/btt559] [Medline: 24072732]
9. Narasimhan B, Rubin DL, Gross SM, Bendersky M, Lavori PW. Software for distributed computation on medical databases: a demonstration project. J Stat Soft 2017;77(13). [doi: 10.18637/jss.v077.i13]
10. OBiBa Opal Documentation. 2019. R DataSHIELD Introduction URL: http://opaldoc.obiba.org/en/latest/r-user-guide/index.html [accessed 2018-02-18]
11. Meeker D, Jiang X, Matheny ME, Farcas C, D'Arcy M, Pearlman L, et al. A system to build distributed multivariate models and manage disparate data sharing policies: implementation in the scalable national network for effectiveness research. J Am Med Inform Assoc 2015 Nov;22(6):1187-1195 [FREE Full text] [doi: 10.1093/jamia/ocv017] [Medline: 26142423]
12. Her QL, Vilk Y, Young J, Zhang Z, Malenfant J, Malek S, et al. 2018 Aug. A Distributed Regression Analysis Application Based on SAS Software. Part I: Linear and Logistic Regression URL: https://ui.adsabs.harvard.edu/#abs/2018arXiv180802387H [accessed 2019-04-15]
13. Vilk Y, Zhang Z, Young J, Her Q, Malenfant J, Malek S, et al. 2018 Aug. A Distributed Regression Analysis Application Based on SAS Software Part II: Cox Proportional Hazards Regression URL: https://ui.adsabs.harvard.edu/#abs/2018arXiv180802392V [accessed 2019-04-15]
14. Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP. Secure analysis of distributed chemical databases without data integration. J Comput Aided Mol Des 2005;19(9-10):739-747. [doi: 10.1007/s10822-005-9011-5] [Medline: 16267693]
15. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016 May;23(3):570-579 [FREE Full text] [doi: 10.1093/jamia/ocv146] [Medline: 26554428]
16. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The US Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf 2012 Jan;21(Suppl 1):1-8. [doi: 10.1002/pds.2343] [Medline: 22262586]
17. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. Clin Pharmacol Ther 2016 Mar;99(3):265-268. [doi: 10.1002/cpt.320] [Medline: 26667601]
18. Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf 2012 Jan;21(Suppl 1):23-31. [doi: 10.1002/pds.2336] [Medline: 22262590]
19. Sentinel System. Routine Querying Tools URL: https://www.sentinelsystem.org/sentinel/surveillance-tools/routine-querying-tools [accessed 2016-02-11]
20. Davies M, Erickson K, Wyner Z, Malenfant J, Rosen R, Brown J. Software-enabled distributed network governance: the PopMedNet experience. EGEMS (Wash DC) 2016;4(2):1213 [FREE Full text] [doi: 10.13063/2327-9214.1213] [Medline: 27141522]
21. Karr AF, Lin X, Sanil AP, Reiter JP. Privacy-preserving analysis of vertically partitioned data using secure matrix products. J Off Stat 2009;25(1):125-138 [FREE Full text]
22. Dankar FK. Privacy preserving linear regression on distributed databases. Trans Data Privacy 2015;8(1):3-28 [FREE Full text]
23. Du W, Han YS, Chen S. Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification. In: Proceedings of the 2004 SIAM International Conference on Data Mining. 2004 Presented at: SIAM'04; April 22-24, 2004; Florida, USA p. 222-233. [doi: 10.1137/1.9781611972740.21]
24. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582 [FREE Full text] [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]
25. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 2013 Dec;20(e2):e226-e231 [FREE Full text] [doi: 10.1136/amiajnl-2013-001926] [Medline: 23956018]

26. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. J Am Med Inform Assoc 2013 May 1;20(3):453-461 [FREE Full text] [doi: 10.1136/amiajnl-2011-000735] [Medline: 22871397]

Abbreviations

AIC: Akaike information criterion
AUC: area under the curve
BIC: Bayesian information criterion
CIDA: Cohort Identification and Descriptive Analysis
DDN: distributed data network
DRA: distribution regression analysis
GLORE: Grid Binary LOgistic Regression
ROC: receiver operating characteristic
SCANNER: SCAlable National Network for Effectiveness Research

Edited by G Eysenbach; submitted 17.06.19; peer-reviewed by W Camelo Castillo, S Wang; comments to author 12.09.19; revised version received 05.11.19; accepted 04.02.20; published 04.06.20.

Please cite as:
Her Q, Malenfant J, Zhang Z, Vilk Y, Young J, Tabano D, Hamilton J, Johnson R, Raebel M, Boudreau D, Toh S
Distributed Regression Analysis Application in Large Distributed Data Networks: Analysis of Precision and Operational Performance
JMIR Med Inform 2020;8(6):e15073
URL: https://medinform.jmir.org/2020/6/e15073
doi:10.2196/15073
PMID:32496200

Original Paper

Using Information Technology to Manage the COVID-19 Pandemic: Development of a Technical Framework Based on Practical Experience in China

Qing Ye[1,2*], MA; Jin Zhou[1*], BA; Hong Wu[2], PhD

[1]Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

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

[*]these authors contributed equally

Corresponding Author:
Hong Wu, PhD
School of Medicine and Health Management
Tongji Medical College
Huazhong University of Science and Technology
13 Hangkong Road, Qiaokou District
Wuhan, 430030
China
Phone: 86 13277942186
Email: wuhong634214924@163.com

Abstract

Background: The coronavirus disease (COVID-19) epidemic poses an enormous challenge to the global health system, and governments have taken active preventive and control measures. The health informatics community in China has actively taken action to leverage health information technologies for epidemic monitoring, detection, early warning, prevention and control, and other tasks.

Objective: The aim of this study was to develop a technical framework to respond to the COVID-19 epidemic from a health informatics perspective.

Methods: In this study, we collected health information technology–related information to understand the actions taken by the health informatics community in China during the COVID-19 outbreak and developed a health information technology framework for epidemic response based on health information technology–related measures and methods.

Results: Based on the framework, we review specific health information technology practices for managing the outbreak in China, describe the highlights of their application in detail, and discuss critical issues to consider when using health information technology. Technologies employed include mobile and web-based services such as Internet hospitals and Wechat, big data analyses (including digital contact tracing through QR codes or epidemic prediction), cloud computing, Internet of things, Artificial Intelligence (including the use of drones, robots, and intelligent diagnoses), 5G telemedicine, and clinical information systems to facilitate clinical management for COVID-19.

Conclusions: Practical experience in China shows that health information technologies play a pivotal role in responding to the COVID-19 epidemic.

(JMIR Med Inform 2020;8(6):e19515) doi:10.2196/19515

KEYWORDS

COVID-19; pandemic; health informatics; health information technology; technical framework; privacy protection

Introduction

The coronavirus disease (COVID-19) epidemic has taken on a global pandemic trend, posing a serious challenge to global health care systems [1]. During the outbreak of COVID-19 in Wuhan, China, the government enacted comprehensive and stringent preventive and control measures to bring the outbreak under control quickly. The health informatics community in China, including clinical informatics, public health informatics, consumer health informatics, and clinical research informatics, has actively taken action to leverage health information technology for epidemic monitoring, detection, early warning, prevention and control, and other tasks [2,3]. The Internet of Things (IoT) has provided platforms such as Worldometer [4] that enable people to access data to monitor the COVID-19 epidemic; integration of big data, such as transportation data and location-based services data, is used to model viral activity and provide guide for health care policy makers [5]; artificial intelligence (AI) and deep learning can enhance the detection and diagnosis of COVID-19 and facilitate the discovery of novel drugs [6]. Clearly, health information technology has played meritorious roles in the battle against COVID-19.

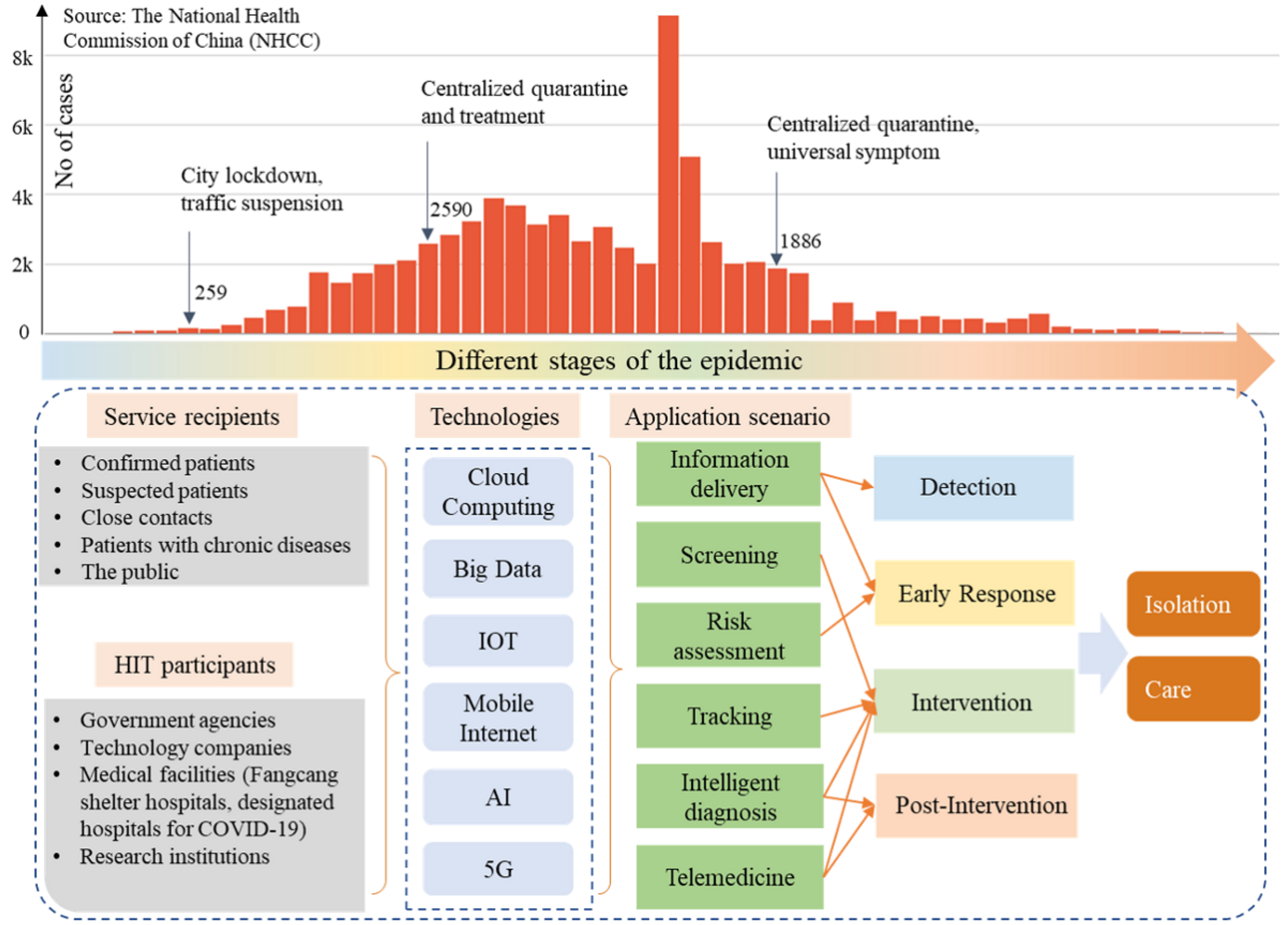
At present, as the COVID-19 epidemic has expanded worldwide, the task of epidemic prevention and control has become more arduous, and the world is facing enormous challenges. To provide theoretical and practical references to other counties on how health information technology can be used to respond to the COVID-19 epidemic as well as to various public health emergencies and disasters, in this study, we develop a technical framework responding to the COVID-19 epidemic from a health informatics perspective. Based on this framework, we also review specific health information technology practices for managing the outbreak in China, describe the highlights of these applications in detail, and finally discuss critical issues to consider when using health information technology.

Methods

Information Gathering and Framework Development

In this study, we collected health information technology–related information released by government departments and management agencies, medical institutions, health care industry associations, and public enterprises to understand the actions taken by the health informatics community in China during the COVID-19 outbreak and to develop a health information technology framework for epidemic response based on health information technology–related measures and methods (Figure 1).

Figure 1. Proposed health information technology framework for responding to the COVID-19 epidemic. COVID-19: coronavirus disease. IoT: Internet of Things.



We first defined and described the health information technology elements of response to an outbreak in terms of health information technology participants, service recipients, technologies, and application scenarios. Second, we constructed a complete technical response solution by correlating the health information technology elements with the different stages of the epidemic. Through this framework, we provide clear understanding of how health information technology is being applied to the epidemic response and how it functions at different stages of the epidemic.

The most effective measures to control infectious diseases are isolation and care; all activities to control the epidemic revolve around these two objectives, as has been demonstrated in China

[7,8]. This framework includes health information technology participants, service recipients, technologies, and application scenarios around the four main stages of the COVID-19 epidemic: detection, early response, intervention, and postintervention.

Definition of the Four Main Stages of the COVID-19 Epidemic

Based on the timeline of the epidemic in China, we identified four main stages (Figure 1): health information technology participants, service recipients, technologies, and application scenarios. This identification is consistent with previous studies [2].

Health Information Technology Participants

Health information technology participants include government agencies, technology companies, medical facilities (Fangcang shelter hospitals, which are designated hospitals for COVID-19), and research institutions, who leverage information technologies to respond to the epidemic in this national campaign [9], were identified.

Service Recipients

These include confirmed patients, suspected patients, close contacts, patients with chronic diseases, and the public [7,10,11].

Technologies

A variety of emerging technologies, such as cloud computing, big data, the IoT, mobile internet, AI, and fifth-generation mobile networks (5G), are being used for epidemic prevention and control. The transition of “5G+ Health” from the experimental to the clinical phase has been realized [12].

Application Scenarios

Application scenarios mainly include information delivery, case detection, screening, online services, risk assessment, and intelligent diagnosis. These scenarios are concrete manifestations of the integrated application of various information technologies for surveillance as well as prevention and control services [11,13-16].

Results

Health Information Technology Practice in China

Information technology has played a key role in China’s response to the COVID-19 outbreak. Information technology was used at all stages of the epidemic, such as prediction of epidemic trends, tracking of close contacts, and remote diagnosis. Based on the health information technology framework for responding to the epidemic described in the Methods section, we present specific health information technology practices for managing the COVID-19 outbreak in China (Table 1) and describe several health information technologies used to fight COVID-19 in detail.

Table 1. Health information technology practices for managing the COVID-19 outbreak in China.

Technology	Scenarios	Application	Examples	References
Mobile internet	Internet hospital, web-based services	Provide a variety of web-based services for the public during the outbreak, including screening and consultation services for mental health disorders or other diseases	Chunyu Yisheng, WeDoctor, China Mobile	Gong et al [11], Liu et al [13], Zheng et al [17], Sun et al [18]
	Web-based information dissemination platforms	Release official statistics about the COVID-19[a] epidemic and keep the public correctly informed about the current situation in a timely fashion	People's Daily, WeChat official account "Healthy China"	People's Daily [19], NHCPRC [20]
Big data	Contact tracing	Record health status and activity trajectory, monitor crowd movement, or locate close contacts	Health QR codes, the Close Contact Detector app	Diao et al [21], Wang et al [22], Boulos et al [23], Ienca et al [2], Boulos et al [23]
	Epidemic prediction	Apply predictive modeling and turning point projection, monitor crowd activity	Predictive model for COVID-19	Wang et al [22], Liu et al [24]
	Spread track	Assist the development of epidemic prevention and control strategies	The dynamic information query system	Zhou et al [25], Peng et al [26]
Cloud computing	Supercomputing	Provide computing power	Supercomputing for big data analytics, vaccine development, and drug development	Ali Group [27], Liu et al [24], Li et al [28]
IoT[b]	Real time data collection	Intelligently manage information	Intelligent Diagnosis and Treatment Assistant Program	Bai et al [29]
AI[c]	Drones	Deploy for fever detection and crowd activity monitoring	DJI drones	Liu et al [24]
	Intelligent diagnosis	Assist doctors in CT[d] diagnosis, reduce work pressure, and improve diagnostic accuracy	Deep learning–based computer-aided diagnostic system	Li et al [30], Gozes et al [15]
	Temperature detection	Rapidly measure body temperature	Airport infrared thermal cameras	Baidu [31]
	Robots	Use intelligent robots to perform simple operations such as disinfection and delivering medications and food during the epidemic	Robots for disinfection, delivering medications, and measuring vital signs	Brickwood [32], Huber [33], Yang et al [34]
5G	5G+ telemedicine	Provide support for remote video consultations and diagnostics	5G telehealth system	Paul [35], Augenstein [36], Li et al [37]
Comprehensive	Clinical information systems	Facilitate clinical management related to COVID-19	Electronic health records	Ren et al [38]

[a]COVID-19: coronavirus disease.

[b]IoT: Internet of Things.

[c]AI: artificial intelligence.

[d]CT: computed tomography.

Internet Hospitals

At present, the global COVID-19 epidemic situation is very serious, and the task of epidemic prevention and control is highly challenging. In the early stages of the epidemic, fever clinics for outpatient and hospital beds were severely overloaded in some areas in China. Against this backdrop, local governments, health care institutions, and a range of companies in China are taking full advantage of mobile internet and 5G technologies to actively provide internet health care services by clinical experts from all over the country. Internet hospitals have played an important role in the prevention and control of epidemics in China [11].

During the COVID-19 outbreak, government agencies encouraged the provision of "Internet+" medical insurance services, which has significantly promoted the public utilization of Internet hospitals [39]. Due to the rapid increase in the number of confirmed cases and deaths during the epidemic, both medical staff and the public have experienced psychological problems, including anxiety and depression. As a result, internet hospitals have started to offer several types of online mental health services [40-42]. For patients with chronic diseases, home delivery services provided by internet hospitals are also in great demand during the outbreak.

Health QR Codes

Health QR codes are a major innovation that have been used during the COVID-19 epidemic for individual tracking at the national level; this has played an important role in epidemic prevention and control as well as in enabling people to return to work [2,23]. People are required to show or scan their health QR code when entering and leaving public places such as communities, supermarkets, and subways. Therefore, a big data system can track the travel routines of an individual based on these records [23]. Health QR codes and big data technology can identify whether a member of the public has been in direct or indirect contact with a confirmed or suspected patient with COVID-19. Through traceability, government agencies can quickly locate potentially infected people and take timely measures to prevent the spread of the virus.

There are currently three colors of health QR codes: red, yellow, and green (Figure 2). These colors indicate three states of health. Different regions have different definitions and requirements for red and yellow codes, while green codes uniformly indicate that a person currently has no symptoms of COVID-19; these codes can be used to quickly judge an individual's health status. Health QR codes not only play a major role in epidemic prevention and control but will also greatly enhance the digital transformation of government and the efficiency of public services.

Figure 2. Health QR codes used in China during the COVID-19 epidemic. COVID-19: coronavirus disease.

Intelligent Diagnosis for Chest Computed Tomography Images

Early in the COVID-19 outbreak, researchers and the clinical informatics sector acted quickly to develop computer-assisted diagnostic products for COVID-19 in collaboration with radiologists [30]. These products have played an important role in screening patients for COVID-19; two of the reasons for this are outlined below.

First, real time reverse transcriptase–polymerase chain reaction (RT-PCR) test results of some patients will show false negative results; as a result, suspected or confirmed patients with COVID-19 are not detected, which is not conducive to disease prevention or control of the epidemic [43,44]. Chest computed tomography (CT) features combined with RT-PCR test results enable more reliable diagnosis in clinical practice. It has been suggested that in addition to RT-PCR results and epidemiological information, special attention should be paid to chest CT features and laboratory examination results [30]. Existing studies show that approximately 96% of patients with COVID-19 present with chest CT abnormalities; therefore, chest CT features are essential for identifying COVID-19 [30].

Second, although RT-PCR serves as the gold standard method for confirmation of COVID-19, the procedure takes a long time (approximately two hours); however, an AI-based diagnostic system can detect lesions of COVID-19 with high sensitivity within two minutes [45]. AI-based diagnosis systems accelerate the screening process for suspected patients, enable the triage of suspected patients in a shorter time, reduce the risk of cross-infection in health care facilities, and relieve the shortage of doctors during the COVID-19 epidemic.

Critical Issues for the Health Informatics Community to Consider

Capabilities of Future Clinical Information Systems

During the COVID-19 epidemic, clinical informatics professionals have also actively participated in the provision of technical support for the admission and treatment of patients with COVID-19. China's experience demonstrates that the design and development of future clinical information systems should focus on the following capabilities to better respond to public health emergencies.

The first capability is rapid deployment. The first three Fangcang shelter hospitals with 4000 beds in Wuhan were built in 29 hours [46]. Clinical information systems must be deployed within the same short time frame to support the admission and treatment of patients with mild to moderate COVID-19.

The second capability is information exchange [47]. In this outbreak, it has frequently been necessary to exchange information (eg, test results and patient referral information) between medical institutions, the Chinese Center for Disease Control and Prevention, the Fangcang shelter hospitals, and designated hospitals for patients with COVID-19. Information exchange capability should be the focus of future clinical information system design to respond to public health emergencies.

The third capability is rapid response of electronic health records (EHRs) to emergencies [3,38]. During this outbreak, clinical informatics professionals have configured the EHR to specifically respond to COVID-19. These configuration and adaptation measures include screen and triage processes, order tools, suspected case reports, and outbreak-related information statistics. A standardized EHR configuration process should be developed to quickly respond to public health emergencies.

Emerging Technologies for Public Health Emergencies

Unlike the severe acute respiratory syndrome (SARS) outbreak in 2003, the internet has become the main information platform during the COVID-19 outbreak, and the public can access dynamic information on the epidemic through various platforms. After nearly 20 years of development, China has made great strides in many emerging technology areas. In the fight against the COVID-19 epidemic, the Chinese government, medical institutions, and a range of technology companies have actively leveraged cloud computing, big data, the IoT, mobile internet, AI, blockchain, 5G, and other digital technologies to improve the efficiency of epidemic monitoring, virus tracking, disease prevention, control, and treatment, resource allocation, etc.

The public can access epidemic situation dynamics and prevention knowledge through the mobile internet. Big data technologies can be used for epidemic situation analysis, material allocation and monitoring of personnel movement. AI has been leveraged in intelligent diagnosis of medical imaging and temperature measurement technology based on computer vision and infrared technology. Telemedicine based on 5G technology has also played an important role in the treatment of patients with severe COVID-19 and in international cooperation in the battle against the outbreak. All these technologies are supported by cloud computing.

Through the Chinese experience, we can clearly see that emerging technologies have played very important roles in epidemic prevention and control; however, we should also note that substantial challenges remain to be faced. While many countries worldwide have well-developed surveillance and prevention systems in place, the health information technology community still has a lot of work to do in light of the global pandemic of COVID-19.

Privacy Protection

Tracing human activity is an important method of identifying the source of COVID-19 infection and controlling the spread of the virus. Individual health information based on health QR codes is an effective measure to monitor and limit the movement of people. In addition, the exchange of health information between different organizations and medical institutions is a basic requirement for guaranteeing patient treatment and care. Chinese practical experience has proven that under the premise of protecting personal privacy, the collection, reasonable use, and exchange of personal information can improve the efficiency and effectiveness of epidemic prevention and control. However, it is also required that personal information be used only with the consent of the person from whom it is collected and that it not be limited to key populations such as confirmed patients, suspected patients, and close contacts.

In the future, the health informatics community should consider establishing a unified framework for the exchange of epidemiological data, subject to privacy protection–related laws, to fully share information and facilitate the orderly flow of information between governments, agencies, and communities to address the challenges posed by public health emergencies and disasters [47].

Discussion

This study makes important theoretical and practical contributions. First, this paper discusses informatics response and experience in responding to the COVID-19 epidemic in China through a health informatics lens. We may have made oversights in our case collection; however, that does not detract from our demonstration of the efforts made by the Chinese health information community and the results they have achieved. Practical experience in China has demonstrated that emerging technologies have unique advantages and can play pivotal roles in addressing major public health challenges. Other countries facing a COVID-19 pandemic should consider using health information technology as part of their public health response. The COVID-19 epidemic is a common challenge facing humanity, and as the epidemic spreads, health information professionals from all countries must share their experiences and work together to explore a complete information technology response framework to improve the response to the current COVID-19 pandemic and to future public health emergencies.

Second, future clinical information systems, personal health records, and big data infrastructures should be capable of rapid adaptation to public health emergencies, such as good interface design and information sharing. If large amounts of personal information are shared, care should be taken to protect user privacy. Moreover, we must consider multiple effects of information technology, such as the spread of “fake news” and rumors, as well as the ethical and privacy issues involved in the leverage of AI and big data technologies [48,49]. Due to the popularity of the Internet and social media, “information epidemics” often occur along with disease outbreaks; therefore, it is also very important to use information technology to increase the transparency of information on epidemics, reduce public panic, and enhance public confidence in the measures taken to combat epidemics [50,51]. During epidemic prevention and control periods, the technology side (technology companies) should prevent technology abuse, while the regulatory side (government agencies and platforms) should be careful to promote technology for the good and benefit of the public.

The COVID-19 epidemic is currently showing a global pandemic trend, and our understanding of the new coronavirus is deepening. Global health information technology practitioners should be proactive and use their professional skills to respond to the COVID-19 epidemic. Practice in China shows that health information technologies play a very important role in responding to the COVID-19 epidemic. Therefore, we believe that the health informatics communities in all countries should react quickly and make full use of health information technology to respond to the epidemic.

Conflicts of Interest

None declared.

References

1. Lenert L, McSwain BY. Balancing Health Privacy, Health Information Exchange and Research in the Context of the COVID-19 Pandemic. J Am Med Inform Assoc 2020 Mar 31 [FREE Full text] [doi: 10.1093/jamia/ocaa039] [Medline: 32232432]
2. Ienca M, Vayena E. On the responsible use of digital data to tackle the COVID-19 pandemic. Nat Med 2020 Apr;26(4):463-464 [FREE Full text] [doi: 10.1038/s41591-020-0832-5] [Medline: 32284619]
3. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid Response to COVID-19: Health Informatics Support for Outbreak Management in an Academic Health System. J Am Med Inform Assoc 2020 Mar 24 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
4. Worldometer. COVID-19 Coronavirus Pandemic URL: https://www.worldometers.info/coronavirus/ [accessed 2020-06-04]
5. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020 Feb;395(10225):689-697 [FREE Full text] [doi: 10.1016/s0140-6736(20)30260-9]
6. Ting DSW, Carin L, Dzau V, Wong TY. Digital technology and COVID-19. Nat Med 2020 Apr;26(4):459-461 [FREE Full text] [doi: 10.1038/s41591-020-0824-5] [Medline: 32284618]
7. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. JAMA 2020 Feb 24. [doi: 10.1001/jama.2020.2648] [Medline: 32091533]
8. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis C, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob Health 2020 Apr;8(4):e488-e496 [FREE Full text] [doi: 10.1016/s2214-109x(20)30074-7]
9. Chen S, Zhang Z, Yang J, Wang J, Zhai X, Bärnighausen T, et al. Fangcang shelter hospitals: a novel concept for responding to public health emergencies. Lancet 2020 Apr;395(10232):1305-1314 [FREE Full text] [doi: 10.1016/s0140-6736(20)30744-3]
10. Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, et al. Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China. JAMA 2020 Apr 10 [FREE Full text] [doi: 10.1001/jama.2020.6130] [Medline: 32275295]
11. Gong K, Xu Z, Cai Z, Chen Y, Wang Z. Internet Hospitals Help Prevent and Control the Epidemic of COVID-19 in China: Multicenter User Profiling Study. J Med Internet Res 2020 Apr 14;22(4):e18908 [FREE Full text] [doi: 10.2196/18908] [Medline: 32250962]
12. Qi X. World Economic Forum. 2020 Apr 08. How next-generation information technologies tackled COVID-19 in China URL: https://www.weforum.org/agenda/2020/04/how-next-generation-information-technologies-tackled-covid-19-in-china/ [accessed 2020-04-14]
13. Liu S, Yang L, Zhang C, Xiang Y, Liu Z, Hu S, et al. Online mental health services in China during the COVID-19 outbreak. Lancet Psychiat 2020 Apr;7(4):e17-e18 [FREE Full text] [doi: 10.1016/s2215-0366(20)30077-8]
14. Hollander JE, Carr BG. Virtually Perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681 [FREE Full text] [doi: 10.1056/nejmp2003539]
15. Adam O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, et al. arXiv. 2020 Mar 10. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis URL: http://arxiv.org/abs/2003.05037 [accessed 2020-04-16]
16. Yang Y, Li W, Zhang Q, Zhang L, Cheung T, Xiang Y. Mental health services for older adults in China during the COVID-19 outbreak. Lancet Psychiat 2020 Apr;7(4):e19 [FREE Full text] [doi: 10.1016/s2215-0366(20)30079-1]
17. Zheng SQ, Yang L, Zhou PX, Li HB, Liu F, Zhao RS. Recommendations and guidance for providing pharmaceutical care services during COVID-19 pandemic: A China perspective. Res Social Adm Pharm 2020 Mar 26 [FREE Full text] [doi: 10.1016/j.sapharm.2020.03.012] [Medline: 32249102]
18. Sun S, Yu K, Xie Z, Pan X. China empowers Internet hospital to fight against COVID-19. J Infection 2020 Apr [FREE Full text] [doi: 10.1016/j.jinf.2020.03.061]
19. People's Daily. Website in Chinese. URL: http://www.people.com.cn/ [accessed 2020-04-17]
20. National Health Commission of the People’s Republic of China. Updates on the epidemic. Website in Chinese URL: http://www.nhc.gov.cn/yjb/s7860/new_list.shtml [accessed 2020-04-17]
21. Diao MY, Zhang S, Chen D, Hu W. The novel coronavirus (COVID-19) infection in Hangzhou: An experience to share. Infect Control Hosp Epidemiol 2020 Mar 05:1-2 [FREE Full text] [doi: 10.1017/ice.2020.62] [Medline: 32131914]
22. Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing. JAMA 2020 Mar 03. [doi: 10.1001/jama.2020.3151] [Medline: 32125371]

23. Kamel Boulos MN, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. Int J Health Geogr 2020 Mar 11;19(1):8 [FREE Full text] [doi: 10.1186/s12942-020-00202-8] [Medline: 32160889]
24. Liu J, HIMSS Greater China Team. Imaging Technology News. Deployment of Health IT in China's Fight Against the COVID-19 Pandemic URL: https://www.itnonline.com/article/deployment-health-it-china%E2%80%99s-fight-against-covid-19-pandemic [accessed 2020-04-17]
25. Zhou C, Su F, Pei T, Zhang A, Du Y, Luo B, et al. COVID-19: Challenges to GIS with Big Data. Geography and Sustainability 2020 Mar;1(1):77-87 [FREE Full text] [doi: 10.1016/j.geosus.2020.03.005]
26. Peng Liu L, Yang W, Zhang D, Zhuge C, Hong L. arXiv. 2020 Feb 16. Epidemic analysis of COVID-19 in China by dynamical modeling URL: https://arxiv.org/abs/2002.06563v1 [accessed 2020-04-17]
27. Alibaba Cloud. Alibaba Cloud Computing Platform URL: https://www.aliyun.com/ [accessed 2020-04-17]
28. Li Z, Li X, Huang YY, Wu Y, Zhou L, Liu R, et al. bioRxiv. 2020 May 25. FEP-based screening prompts drug repositioning against COVID-19 URL: https://www.biorxiv.org/content/10.1101/2020.03.23.004580v1 [accessed 2020-06-04]
29. Bai L, Yang D, Wang X, Tong L, Zhu X, Zhong N, et al. Chinese experts' consensus on the Internet of Things-aided diagnosis and treatment of coronavirus disease 2019 (COVID-19). Clinical eHealth 2020;3:7-15 [FREE Full text] [doi: 10.1016/j.ceh.2020.03.001]
30. Li D, Wang D, Dong J, Wang N, Huang H, Xu H, et al. False-Negative Results of Real-Time Reverse-Transcriptase Polymerase Chain Reaction for Severe Acute Respiratory Syndrome Coronavirus 2: Role of Deep-Learning-Based CT Diagnosis and Insights from Two Cases. Korean J Radiol 2020 Apr;21(4):505-508 [FREE Full text] [doi: 10.3348/kjr.2020.0146] [Medline: 32174053]
31. Baidu. 2020 Feb 02. AI temperature detection technology landing in Beijing. Article in Chinese URL: https://baijiahao.baidu.com/s?id=1657416555899875487&wfr=spider&for=pc [accessed 2020-04-17]
32. Health Europa. 2020 Feb 05. XAG introduces drone disinfection operation to fight the coronavirus outbreak URL: https://www.healtheuropa.eu/xag-introduces-drone-disinfection-operation-to-fight-the-coronavirus-outbreak/97265/ [accessed 2020-04-16]
33. Huber M. Aviation International News. 2020 Feb 07. Drones Enlisted To Fight Corona Virus in China URL: https://www.ainonline.com/aviation-news/general-aviation/2020-02-07/drones-enlisted-fight-corona-virus-china [accessed 2020-04-16]
34. Yang G, J. Nelson B, Murphy RR, Choset H, Christensen H, H. Collins S, et al. Combating COVID-19—The role of robotics in managing public health and infectious diseases. Sci Robot 2020 Mar 25;5(40):eabb5589 [FREE Full text] [doi: 10.1126/scirobotics.abb5589]
35. Paul G. Business Insider. 2020 Jan 28. ZTE and China Telecom enabled the first remote diagnosis of coronavirus via a 5G telehealth system URL: https://www.businessinsider.com/zte-china-telecom-build-5g-telehealth-system-for-coronavirus-2020-1?international=true&r=US&IR=T [accessed 2020-04-17]
36. Augenstein J. Health Affairs. 2020 Mar 16. Opportunities to Expand Telehealth Use amid the Coronavirus Pandemic URL: https://www.healthaffairs.org/do/10.1377/hblog20200315.319008/full/ [accessed 2020-04-17]
37. Li H, Zheng S, Liu F, Liu W, Zhao R. Fighting against COVID-19: Innovative strategies for clinical pharmacists. Res Social Adm Pharm 2020 Apr 06 [FREE Full text] [doi: 10.1016/j.sapharm.2020.04.003] [Medline: 32278766]
38. Ren Y, Zhang X, Li J, Wang Y, San J, Nai C. Tongji Medical College Huazhong University of Science & Technology experiences on novel coronavirus epidemic prevention and treatment information system support. Chin J Hosp Admin 2020;36:E003. [doi: 10.3760/cma.j.issn.1000-6672.2020.0003]
39. National Health Insurance Agency. 2020 Feb 28. Guidance on Promoting "Internet+" Medical Insurance Services during the Prevention and Control of COVID-19 Epidemic. Article in Chinese URL: http://www.gov.cn/zhengce/zhengceku/2020-03/03/content_5486256.htm [accessed 2020-04-14]
40. Kang L, Li Y, Hu S, Chen M, Yang C, Yang BX, et al. The mental health of medical workers in Wuhan, China dealing with the 2019 novel coronavirus. Lancet Psychiat 2020 Mar;7(3):e14 [FREE Full text] [doi: 10.1016/s2215-0366(20)30047-x]
41. Xiang Y, Yang Y, Li W, Zhang L, Zhang Q, Cheung T, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. Lancet Psychiat 2020 Mar;7(3):228-229 [FREE Full text] [doi: 10.1016/s2215-0366(20)30046-8]
42. Duan L, Zhu G. Psychological interventions for people affected by the COVID-19 epidemic. Lancet Psychiat 2020 Apr;7(4):300-302 [FREE Full text] [doi: 10.1016/s2215-0366(20)30073-0]
43. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing. Radiology 2020 Feb 12:200343. [doi: 10.1148/radiol.2020200343] [Medline: 32049601]
44. Lan L, Xu D, Ye G, Xia C, Wang S, Li Y, et al. Positive RT-PCR Test Results in Patients Recovered From COVID-19. JAMA 2020 Feb 27 [FREE Full text] [doi: 10.1001/jama.2020.2783] [Medline: 32105304]
45. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. Radiology 2020 Feb 19:200432. [doi: 10.1148/radiol.2020200432] [Medline: 32073353]
46. Chen S, Zhang Z, Yang J, Wang J, Zhai X, Bärnighausen T, et al. Fangcang shelter hospitals: a novel concept for responding to public health emergencies. Lancet 2020 Apr;395(10232):1305-1314 [FREE Full text] [doi: 10.1016/s0140-6736(20)30744-3]

47. Lenert L, McSwain BY. Balancing Health Privacy, Health Information Exchange and Research in the Context of the COVID-19 Pandemic. J Am Med Inform Assoc 2020 Mar 31 [FREE Full text] [doi: 10.1093/jamia/ocaa039] [Medline: 32232432]
48. Wong JEL, Leo YS, Tan CC. COVID-19 in Singapore-Current Experience: Critical Global Issues That Require Attention and Action. JAMA 2020 Feb 20. [doi: 10.1001/jama.2020.2467] [Medline: 32077901]
49. Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H. The pandemic of social media panic travels faster than the COVID-19 outbreak. J Travel Med 2020 May 18;27(3) [FREE Full text] [doi: 10.1093/jtm/taaa031] [Medline: 32125413]
50. Kandhway K, Kuri J. Campaigning in Heterogeneous Social Networks: Optimal Control of SI Information Epidemics. IEEE/ACM Trans Networking 2016 Feb;24(1):383-396 [FREE Full text] [doi: 10.1109/tnet.2014.2361801]
51. Chen E, Lerman K, Ferrara E. arXiv. 2020 Jun 02. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set URL: http://arxiv.org/abs/2003.07372 [accessed 2020-06-04]

Abbreviations

AI: artificial intelligence
COVID-19: coronavirus disease
CT: computerized tomography
EHR: electronic health record
IoT: Internet of Things
RT-PCR: reverse transcriptase–polymerase chain reaction
SARS: severe acute respiratory syndrome

Edited by G Eysenbach; submitted 21.04.20; peer-reviewed by C Zhang, W Song, H Yu; comments to author 27.05.20; revised version received 29.05.20; accepted 30.05.20; published 08.06.20.

Please cite as:
Ye Q, Zhou J, Wu H
Using Information Technology to Manage the COVID-19 Pandemic: Development of a Technical Framework Based on Practical Experience in China
JMIR Med Inform 2020;8(6):e19515
URL: http://medinform.jmir.org/2020/6/e19515/
doi:10.2196/19515
PMID:32479411

Original Paper

Identification of the Best Semantic Expansion to Query PubMed Through Automatic Performance Assessment of Four Search Strategies on All Medical Subject Heading Descriptors: Comparative Study

Clément R Massonnaud[1,2], MPH; Gaétan Kerdelhué[1,2], MSc; Julien Grosjean[1,2], PhD; Romain Lelong[1,2], PhD; Nicolas Griffon[1,2], PhD; Stefan J Darmoni[1,2], MD, PhD

[1]Department of Biomedical Informatics, Rouen University Hospital, Rouen, France

[2]Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, U1142, INSERM, Sorbonne Université, Paris, France

Corresponding Author:
Clément R Massonnaud, MPH
Department of Biomedical Informatics
Rouen University Hospital
1 rue de Germont
Rouen
France
Phone: 33 2 32 88 89 90
Email: clement.massonnaud@gmail.com

Abstract

Background: With the continuous expansion of available biomedical data, efficient and effective information retrieval has become of utmost importance. Semantic expansion of queries using synonyms may improve information retrieval.

Objective: The aim of this study was to automatically construct and evaluate expanded PubMed queries of the form *“preferred term”[MH] OR “preferred term”[TIAB] OR “synonym 1”[TIAB] OR “synonym 2”[TIAB] OR …*, for each of the 28,313 Medical Subject Heading (MeSH) descriptors, by using different semantic expansion strategies. We sought to propose an innovative method that could automatically evaluate these strategies, based on the three main metrics used in information science (precision, recall, and F-measure).

Methods: Three semantic expansion strategies were assessed. They differed by the synonyms used to build the queries as follows: MeSH synonyms, Unified Medical Language System (UMLS) mappings, and custom mappings (Catalogue et Index des Sites Médicaux de langue Française [CISMeF]). The precision, recall, and F-measure metrics were automatically computed for the three strategies and for the standard automatic term mapping (ATM) of PubMed. The method to automatically compute the metrics involved computing the number of all relevant citations (A), using National Library of Medicine indexing as the gold standard (*“preferred term”[MH]*), the number of citations retrieved by the added terms (*”synonym 1“[TIAB] OR ”synonym 2“[TIAB] OR …*) (B), and the number of relevant citations retrieved by the added terms (combining the previous two queries with an “AND” operator) (C). It was possible to programmatically compute the metrics for each strategy using each of the 28,313 MeSH descriptors as a “preferred term,” corresponding to 239,724 different queries built and sent to the PubMed application program interface. The four search strategies were ranked and compared for each metric.

Results: ATM had the worst performance for all three metrics among the four strategies. The MeSH strategy had the best mean precision (51%, SD 23%). The UMLS strategy had the best recall and F-measure (41%, SD 31% and 36%, SD 24%, respectively). CISMeF had the second best recall and F-measure (40%, SD 31% and 35%, SD 24%, respectively). However, considering a cutoff of 5%, CISMeF had better precision than UMLS for 1180 descriptors, better recall for 793 descriptors, and better F-measure for 678 descriptors.

Conclusions: This study highlights the importance of using semantic expansion strategies to improve information retrieval. However, the performances of a given strategy, relatively to another, varied greatly depending on the MeSH descriptor. These results confirm there is no ideal search strategy for all descriptors. Different semantic expansions should be used depending on the descriptor and the user’s objectives. Thus, we developed an interface that allows users to input a descriptor and then proposes the best semantic expansion to maximize the three main metrics (precision, recall, and F-measure).

(JMIR Med Inform 2020;8(6):e12799) doi:10.2196/12799

KEYWORDS

bibliographic database; information retrieval; literature search; Medical Subject Headings; MEDLINE; PubMed; precision; recall; search strategy; thesaurus

Introduction

Background

With the continuous expansion of available biomedical data, efficient and effective information retrieval has become extremely important. The number of citations for biomedical literature accessible through the PubMed search engine, a service of the US National Library of Medicine (NLM), was over 30 million in January 2019, and of these, more than 26 million were from the MEDLINE database. The number of citations added to MEDLINE each year now exceeds 1 million (1,178,360 citations added in 2016) [1]. PubMed is one of the most used tools to access these data, and its popularity is growing steadily each year (from 2.5 billion searches performed in 2013 to 3.3 billion in 2017) [2].

However, numerous studies have reported that users lack search skills for the effective use of PubMed [3-5]. Although a basic search using PubMed can be relatively straightforward, a deeper understanding of its structure and underlying search algorithm is needed to perform an effective search of the literature. In order to improve the accuracy of information retrieval, MEDLINE citations are indexed in the Medical Subject Headings (MeSH) thesaurus [6], but most users do not know it well enough and do not commonly use its descriptors to build their queries [7,8]. The MeSH thesaurus, developed by the NLM, is a list of descriptors covering the biomedical field. Moreover, users rarely employ search tags and therefore do not fully exploit the features of PubMed [9]. Consequently, the NLM has developed an automatic process to modify users' explicit queries called automatic term mapping (ATM). Entry terms are mapped to their corresponding MeSH descriptors and compound words are broken down and combined with the Boolean operators "AND" and "OR" and searched with the tag [All fields] [10].

Comprehensive literature searching requires the use of both bibliographic database searching and diverse supplementary search methods [11,12]. The purpose of ATM is to improve information retrieval in bibliographic database searching, but several studies have proposed alternative processes to enhance users' queries that have yielded better results. For instance, Kim et al proposed the use of semantic techniques, such as document similarity [13]. PubMed has also implemented the recommendation of a "related articles" feature [14]. This feature uses the PubMed-related citations algorithm, which is a probabilistic topic-based model developed by Lin and Wilbur [15]. Wei et al proposed a strategy to enhance this feature [16]. Additionally, Afzal et al proposed methods for the automation of query generation [17,18]. Other popular strategies propose to perform semantic expansion with synonyms of the entry terms. These strategies vary in the knowledge organization system (KOS) they use to perform the expansions. Aronson et al in 1997 [19] and Hersh et al in 2000 [20] proposed the use of Unified Medical Language System (UMLS) mappings to perform query expansion. The UMLS thesaurus maps terms of different KOSs using concept unique identifiers (CUIs). In 2009, Thirion et al proposed the expansion of queries with MeSH synonyms [21]. This strategy was also explored in 2016 by Wright et al [22]. In the MeSH thesaurus, each descriptor has a preferred term and may have some synonyms. This optimization led to a great improvement in the performance of information retrieval. In 2012, Griffon et al proposed the use of the UMLS to perform the expansion [23], leading to a slight increase in recall but a decrease in precision. Among other strategies, Xu et al [24] proposed a biomedical query expansion framework based on learning-to-rank methods, in which term-ranking models are trained to refine the candidate expansion terms by selecting the most relevant terms for enriching the original query.

Prior Work

In order to improve information retrieval, our team (physicians, librarians, and terminology specialists) developed a new strategy of semantic expansion using new mappings between various KOSs, which was called CISMeF (French acronym, Catalogue et Index des Sites Médicaux de langue Française) mappings. Health Terminology/Ontology Portal (HeTOP) is a cross-lingual multi-terminology server also developed by the CISMeF team, which contains 86 KOSs in 45 languages. The CISMeF mappings are mappings between terms of these 86 KOSs. The mappings were created in various ways. Some concepts were mapped automatically, using UMLS CUIs. However, most of the KOSs included in HeTOP are not included in the UMLS (65 out of the 86 KOSs). For these KOSs, the terms were mapped automatically using natural language processing or manually by librarians and KOS specialists. Moreover, some of the automatically mapped terms were verified and manually curated by librarians and KOS specialists (supervised mappings). Information on HeTOP and CISMeF mappings has been detailed in a previous paper [25]. The performance of this new kind of semantic expansion using CISMeF mappings was manually assessed in a previous study by Massonnaud et al [26].

Although these different strategies have greatly improved the effectiveness of information retrieval, there are limitations to the assessment of their performances. Assessments were performed manually, allowing only small samples of descriptors and citations. Moreover, all the studies revealed a great variability in results depending on the descriptor used. This behavior suggests that there is no semantic expansion that would be optimal for all descriptors and that the semantic expansion to be used should be chosen according to the specific descriptor and the user's objective (ie, when seeking either better precision or recall or a harmonic mean view using F-measure).

Objective

The aim of this study was to automatically construct and evaluate expanded PubMed queries of the form *"preferred* *term"[MH] OR "preferred term"[TIAB] OR "synonym 1"[TIAB] OR "synonym 2"[TIAB] OR…*, for each of the 28,313 MeSH descriptors by using different semantic expansion strategies. We sought to propose an innovative method that could automatically evaluate these strategies, based on the three main metrics used in information science (precision, recall, and F-measure).

Methods

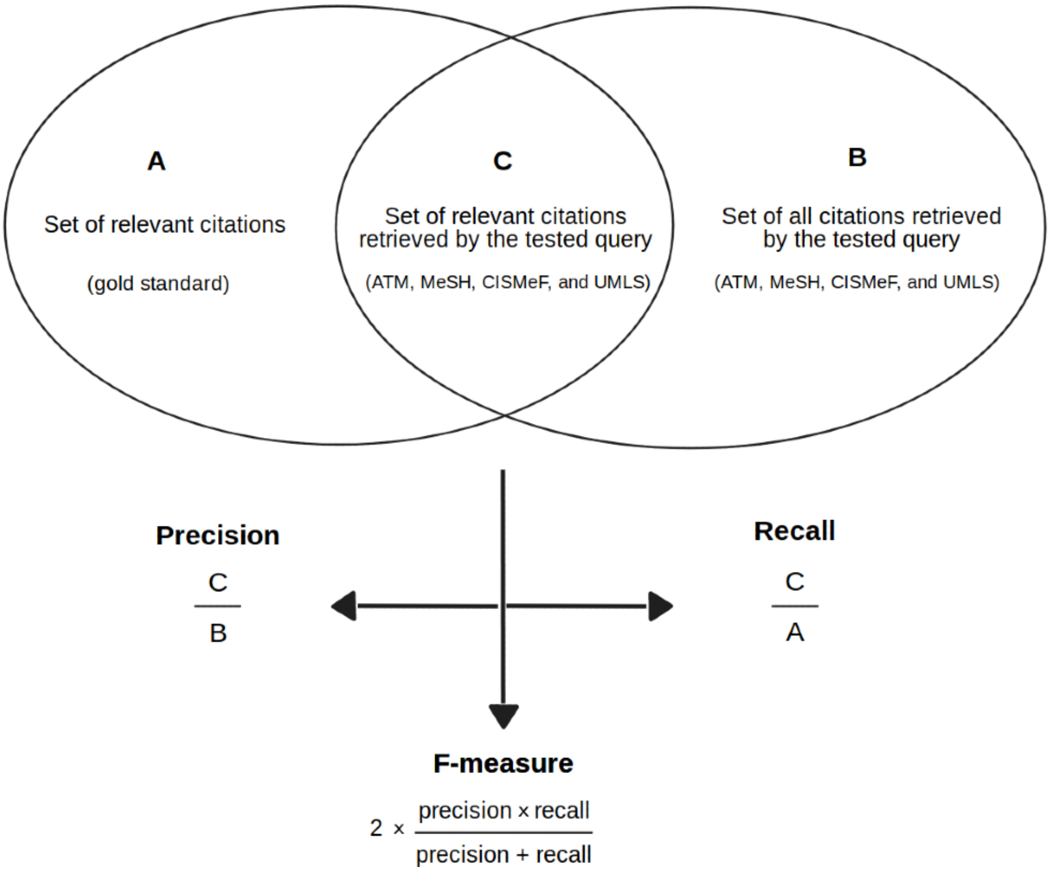
Semantic Expansion Strategies

Four strategies were assessed in this study, including the standard ATM of PubMed and three kinds of queries enhanced with semantic expansions (MeSH, UMLS, and CISMeF). The three semantic expansion strategies differed in the set of synonyms used to expand the query. For example, for the MeSH descriptor ”diabetes mellitus, type 2“[MH], the query built with the ATM strategy was as follows: *(”diabetes mellitus, type 2“[MeSH Terms] OR ”type 2 diabetes mellitus“[All Fields] OR ”diabetes mellitus, type 2“[All Fields])*, and the query built with the UMLS strategy was as follows: *(”diabetes mellitus, type 2“[MH] OR ”adult onset diabetes“[TIAB] OR ”adult onset diabetes mellitus“[TIAB] OR ”adult-onset diabetes“[TIAB] OR ”adult-onset diabetes mellitus“[TIAB] OR …)*. The four strategies were applied for each of the 28,313 MeSH descriptors, and their respective performance was assessed by computing standard metrics (precision, recall, and F-measure).

Automatic Metrics Computation

Figure 1 depicts how the metrics were computed. Precision was defined as the fraction of relevant citations among the retrieved citations. Recall was defined as the fraction of relevant citations retrieved from the total number of relevant citations. The traditional F-measure (or F1 score) was defined as the harmonic mean of the precision and the recall, and the value is provided by the following formula: 2 × (precision × recall) / (precision + recall). Therefore, in order to automatically estimate these metrics for a given query and a given descriptor, it was necessary to identify the set of all relevant citations for the descriptor (A; ie, the gold standard), the set of all citations retrieved by the query (B), and the set of citations retrieved by the query that were relevant (C; ie, the intersection of A and B).

Figure 1. Representation of the three sets of citations retrieved for each descriptor, and how they were used to compute the metrics. ATM: automatic term mapping; CISMeF: Catalogue et Index des Sites Médicaux de langue Française; MeSH: Medical Subject Headings; UMLS: Unified Medical Language System.


A
Set of relevant citations
(gold standard)
C
Set of relevant citations retrieved by the tested query
(ATM, MeSH, CISMeF, and UMLS)
B
Set of all citations retrieved by the tested query
(ATM, MeSH, CISMeF, and UMLS)
Precision
C / B
Recall
C / A
F-measure
2 × precision × recall / precision + recall

The set of relevant citations (A) was defined using NLM's indexing as the gold standard. For a query built from a particular MeSH descriptor, a citation was considered relevant if it was indexed with that same descriptor. Therefore, the total number of relevant citations (A) was retrieved via the following query: *"preferred term"[MH]*. The queries for set (B) were constructed as follows: for the ATM strategy, the query was constructed with *"preferred term"[TIAB]*. For the other three strategies, the query was constructed with the synonyms retrieved by the expansion strategy (*"synonym 1"[TIAB] OR "synonym 2"[TIAB] OR etc.*). Consequently, the number of relevant citations retrieved (C) could be computed by combining the previous two queries with an “AND” operator as follows: *(“preferred term”[MH] AND “preferred term”[TIAB])* for ATM and *(“preferred term”[MH] AND (“synonym 1”[TIAB] OR “synonym 2”[TIAB] OR etc.))* for the three other strategies. The tag *[All fields]* was replaced with *[TIAB]* since *[All fields]* also searches the indexation field of the citations, therefore conflicting with the *[MH]* tag. Moreover, the scope of the search was reduced to indexed citations by adding the tag *medline[sb]* so that all queries were performed on the same set of manually indexed citations. Table 1 shows a summary of the syntax of the resulting nine different queries.

Table 1. Summary of the syntax of the nine different queries used in this study.

Strategy	Relevant citations (A)	Retrieved citations (B)	Relevant citations retrieved (C)
ATM[a]	”pref. term“[MH]	”pref. term“[TIAB] AND medline[sb]	”pref. term“[MH] AND (”pref. term“[TIAB]) AND medline[sb]
MeSH[b]	”pref. term“[MH]	(”MeSH synonym 1“[TIAB] OR ”MeSH synonym 2“[TIAB] OR …) AND medline[sb]	”pref. term“[MH] AND (”MeSH synonym 1“[TIAB] OR ”MeSH synonym 2“[TIAB] OR …) AND medline[sb]
UMLS[c]	”pref. term“[MH]	(”UMLS synonym 1“[TIAB] OR ”UMLS synonym 2“[TIAB] OR …) AND medline[sb]	”pref. term“[MH] AND (”UMLS synonym 1“[TIAB] OR ”UMLS synonym 2“[TIAB] OR …) AND medline[sb]
CISMeF[d]	”pref. term“[MH]	(”CISMeF synonym 1“[TIAB] OR ”CISMeF synonym 2“[TIAB] OR …) AND medline[sb]	”pref. term“[MH] AND (”CISMeF synonym 1“[TIAB] OR ”CISMeF synonym 2“[TIAB] OR …) AND medline[sb]

[a] ATM: automatic term mapping.

[b] MeSH: Medical Subject Headings.

[c] UMLS: Unified Medical Language System.

[d] CISMeF: Catalogue et Index des Sites Médicaux de langue Française.

Data Collection

Initially, the queries were built using the HeTOP terminology server [25], which provides relations between multiple KOSs. Given a particular concept, it is possible to automatically gather the MeSH preferred term of this concept and its synonyms from the KOS of interest. As the 2018 version of the MeSH was not released at the time of this study, the 2017 version containing 28,313 descriptors was used, and of these descriptors, 26,636 were used at least once for indexing citations. It was then possible to programmatically build the nine different types of queries (Table 1) for each of the 26,636 descriptors, resulting in a total of 239,724 queries. ATM’s behavior regarding the split of compound words was reproduced exactly. In order to shorten the length of the queries, the terms were set to lowercase and multiple occurrences of the exact same term were removed. Thereafter, the citation count for each of the 239,724 queries was retrieved via PubMed’s application programming interface. The processing time of these 239,724 queries on a microcomputer was around 3 hours and 30 minutes. Therefore, it is scalable and can be run frequently.

Statistical Analysis

The mean precision, recall, and F-measure were computed for the 26,636 descriptors and for each of the four search strategies. The four strategies were ranked, and the number of descriptors for which the CISMeF strategy had better results than each of the three other strategies was computed, considering a difference of at least 5% (arbitrary). The metrics were also computed with stratification according to the MeSH category. Statistical analysis was performed using R software (version 3.4.3; R Foundation for Statistical Computing, Vienna, Austria). As the analysis was performed on the entire set of MeSH descriptors, confidence intervals and *P* values were not needed and therefore not computed.

Results

Table 2 shows the mean precision, recall, and F-measure for each of the four search strategies. ATM had the worst performance for all three metrics among the four strategies. MeSH had the best mean precision (51%, SD 23%). CISMeF and UMLS had identical results for precision. UMLS had the best recall and F-measure (41% and 36%, respectively). CISMeF had the second best recall and F-measure.

Table 3 shows the number of descriptors for which two strategies had equal precision, recall, or F-measure. Table 4 shows the number of descriptors for which the metric score of a strategy was at least 5% better than another strategy.

Table 2. Mean performances of the four search strategies for the 26,636 Medical Subject Heading descriptors.

KOS[a]	Precision (%), mean (SD)	Recall (%), mean (SD)	F-measure (%), mean (SD)
ATM[b]	44 (24)	31 (29)	28 (23)
MeSH[c]	51 (23)	38 (31)	35 (24)
CISMeF[d]	49 (23)	40 (31)	35 (24)
UMLS[e]	49 (23)	46 (31)	36 (24)

[a]KOS: knowledge organization system.

[b]ATM: automatic term mapping.

[c]MeSH: Medical Subject Headings.

[d]CISMeF: Catalogue et Index des Sites Médicaux de langue Française.

[e]UMLS: Unified Medical Language System.

Table 3. Number of descriptors for which two strategies had equal precision, recall, or F-measure.

KOS[a]	Precision, n	Recall, n	F-measure, n
ATM[b] and MeSH[c]	3037	3959	2938
ATM and CISMeF[d]	2551	3410	2459
ATM and UMLS[e]	2409	3265	2320
MeSH and CISMeF	19,261	20,232	19,001
MeSH and UMLS	17,176	18,394	16,917
CISMeF and UMLS	18,819	19,956	18,565

[a]KOS: knowledge organization system.

[b]ATM: automatic term mapping.

[c]MeSH: Medical Subject Headings.

[d]CISMeF: Catalogue et Index des Sites Médicaux de langue Française.

[e]UMLS: Unified Medical Language System.

Table 4. Comparisons of the strategies for each metric.

Strategy comparison	Metric, n[a]		
	Precision	Recall	F-measure
CISMeF[b] vs UMLS[c]	1180 vs 1017	793 vs 1262	678 vs 1140
MeSH[d] vs UMLS	2285 vs 215	9 vs 2857	553 vs 1761
CISMeF vs MeSH	170 vs 2088	2372 vs 9	1403 vs 669
MeSH vs ATM[e]	9650 vs 3299	8198 vs 3404	8150 vs 2446
CISMeF vs ATM	9112 vs 4557	9724 vs 2949	8895 vs 2628
ATM vs UMLS	4682 vs 9094	2852 vs 10,047	2448 vs 9217

[a]The numbers are the numbers of descriptors for which the metric score of a strategy was at least 5% better than another strategy.

[b]CISMeF: Catalogue et Index des Sites Médicaux de langue Française.

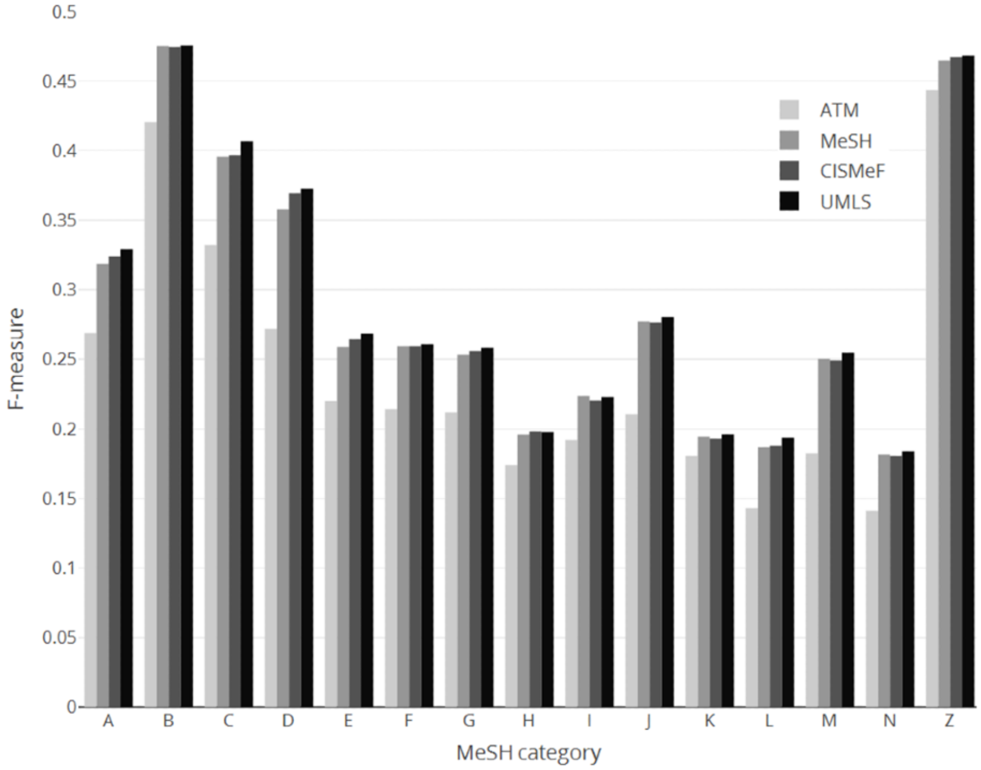
[c]UMLS: Unified Medical Language System.

[d]MeSH: Medical Subject Headings.

[e]ATM: automatic term mapping.

The analysis stratified according to the category (tree) of the MeSH descriptor revealed the same trends for all three metrics. The best precision was obtained with category C (diseases) by the MeSH strategy (58.16%). MeSH had the best precision for all categories expect category B (organisms), for which ATM had the best precision (data not shown). The best recall was obtained with category B by UMLS (64.28%), which had the best results for 11 out of the 15 categories. CISMeF had the best recall for the following remaining four categories: H (disciplines and occupations), K (humanities), L (information science), and N (health care). ATM had the worst recall for all categories (data not shown). Figure 2 shows the F-measure scores for each of the four strategies depending on the MeSH category of the descriptor. The best F-measure was obtained with category B by UMLS (47.55%). UMLS had the best F-measure for all categories except category H, for which CISMeF had the best score (19.82%), and category I (anthropology, education, sociology, and social phenomena), for which MeSH had the best F-measure (22.38%).

Figure 2. F-measure scores of the four search strategies depending on the MeSH category of the descriptor. ATM: automatic term mapping; CISMeF: Catalogue et Index des Sites Médicaux de langue Française; MeSH: Medical Subject Headings; UMLS: Unified Medical Language System.



Discussion

Principal Findings

Of the four strategies assessed in this study, PubMed's standard ATM had the worst mean performances for the three metrics measured (ie, precision, recall, and F-measure). These results are consistent with the findings of previous studies [21,23,26]. The best precision was obtained with the MeSH strategy (50.93%). The mean values of precision for both the CISMeF and UMLS strategies were identical (49.20%). For recall and F-measure, the best performance was obtained by the UMLS strategy, followed by the CISMeF strategy and MeSH strategy.

Even though the differences between the mean performances of the three enhanced strategies (ie, MeSH, CISMeF, and UMLS) were small, with no difference at all for numerous descriptors, the finding did not reflect the important variability of the results. Indeed, for each metric, the ranking of the four strategies was greatly dependent on the descriptor. For instance, although UMLS and CISMeF had identical performances for mean precision, CISMeF had better precision than UMLS for 1180 descriptors. Likewise, the CISMeF strategy had better recall for 793 descriptors and better F-measure for 678 descriptors. Even the ATM strategy, which had much lower mean results for all three metrics, was ranked first for several descriptors (data not shown).

The important variability of the performances found here is consistent with the results of previous studies [21,23,26]. This variability was an important limiting factor for these studies since the assessments were performed manually and therefore on a restricted set of descriptors. Consequently, the interpretation of the results was difficult and had important limitations. The objective of this study was to implement and evaluate an original approach for the automatic assessment of the three main metrics of information science. This innovative method allowed us to test the four different strategies of semantic expansion on the entire set of MeSH preferred descriptors (n=28,313) rather than on a small subset. The results found with this new method conform to those of previous studies, with a similar ranking of the four search strategies. Working out the reasons for variability would be a complex endeavor, as they differ from one descriptor to another. The first obvious reason for a loss of precision, for instance, is an increased number of synonyms used. A higher number of terms in a query is associated with a higher likelihood of it having less precision. For example, the CISMeF expansion for the term "kidney tubular necrosis, acute" had a precision of 97.4% as against 36.6% for the UMLS expansion. The CISMeF expansion included eight terms, whereas the UMLS expansion included 18 terms. Even with a comparable number of terms, a loss in precision could be caused by a single term in the query, for instance, an acronym or a broader synonym (hyperonym). For example, the CISMeF expansion for the term "drug interactions" had a precision of 56.3% as against 5.1% for the

UMLS expansion. Both expansions are based on slight variations of the combination of “drug” and “interactions,” but one term in the UMLS expansion is simply “interactions,” which leads to a lot of noise and thus a decrease in the precision of the query. For recall, the reasons for variation are somewhat reciprocal. For instance, the CISMeF expansion for the term “chronic disease” yields a recall of 73.8% as against 7.8% for the UMLS expansion. Both expansions use variations of the combination of “chronic” and “disease,” but the CISMeF expansion uses an additional “chronic” term alone. However, this gain in recall is at the cost of a loss in precision (13.1% loss).

The ranking of the four strategies was similar after stratification according to the descriptor’s category in the MeSH tree. The exact same evaluation was performed with different tags in the queries [14]. The assessment was also performed over different time intervals. The tag **[majr]* was tested instead of **[mh]*, and all strategies were tested with and without the explosion behavior. The explosion is activated by default in PubMed’s ATM, but it was not feasible to reproduce the explosion in our expanded queries, as this would have resulted in too large queries. This could slightly bias the results for broad MeSH descriptors, but the analysis we performed by deactivating the explosion for the gold standard (using *[mesh:noexp]*) revealed similar ranking of the four strategies and similar variability (data not shown). Moreover, a new version of PubMed is available since November 2019 and provides a modified version of the ATM. Unfortunately, the details about its modifications have not been published at the time of this study, and therefore, it is not reproducible [27]. Future research could analyze the new version of the ATM in comparison with the expansion strategies and with the legacy ATM.

Our results suggest that the choice of the semantic expansion strategy used to build the query must be made according to the descriptor. Since the automatic assessment tested here allowed assessment of all the MeSH preferred descriptors, it is now possible to choose which semantic expansion strategy to use to build a query for a given descriptor, according to the performances of the three metrics precision, recall, and F-measure. As the processing time of this automatic assessment is quite low, it can be updated frequently (each day, each week, or each month). Technically, the assessment could also be performed in real time, although this does not seem necessary since the results should not vary greatly during short periods of time.

The availability of quantitative measurements of the performances of different strategies now allows users to decide which semantic expansion to use given a particular MeSH descriptor. Depending on their specific needs, users could either choose the strategy providing the best precision, best recall, or best F-measure, since these performances could be accomplished by different strategies. These considerations led our team to develop an interface that allows users to input their MeSH preferred descriptor and to choose which metric they wish to maximize (precision, recall, or F-measure). This tool is freely available on the HeTOP website [28] (Multimedia Appendix 1). Users can search for a term, and the algorithm will try to match the term with the corresponding MeSH descriptor. Once on the description page of a descriptor, the “PubMed/Doc’CISMef” tab needs to be selected to go to the query building section. There, in section 2 “options,” users can select which metric they want to maximize (precision, recall, or F-measure). The algorithm will determine which expansion strategy yields the best score for the metric and automatically construct the query with the corresponding semantic expansion. Thereafter, in section 3 “queries,” a button with a PubMed icon appears with a tag corresponding to the semantic expansion that will be used to build the query. By clicking this button, the query will be automatically built and sent to the PubMed search engine in a new browser page. A perspective could be to go even further in the customization of the queries, with the possibility to add each synonym successively, each time assessing in real time the performance of this custom query.

Limitations

This study has some limitations. First, in order to assess the metrics in an automatic manner, the scope of the search had to be restricted to the indexed citations of the MEDLINE database. The assessment of recent nonindexed citations could only be performed manually, with all the limiting factors previously described in the literature. However, it is legitimate to assume that the different semantic expansions would perform in the same way for the entire database since there is no reason to think that the indexing paradigms would shift suddenly for a given descriptor. Moreover, the results presented here are consistent with manual evaluations of previous studies, suggesting there is no major bias in this new methodology. Second, the queries built were simple queries based on only one MeSH preferred term. It would be necessary to evaluate the performances of these different semantic expansions with more complex queries, associating multiple MeSH preferred terms. However, the behavior of such queries would be identical because the semantic expansion of each term would be treated independently and then recombined with Boolean operators, which is the default behavior of PubMed’s ATM.

Conclusions

In this study, we present an innovative method to automatically compute for PubMed citations the three main metrics used in information science. This new method allowed us to compare four semantic expansion strategies to query PubMed on all MeSH descriptors. The results confirmed great variability depending on the descriptor. Hence, there is a need to propose to users the semantic expansion that best fits their specific objectives. Owing to the possibility of regularly updating the performances of these search strategies for all MeSH descriptors, our team has developed an interface that allows users to input a descriptor and then proposes the best semantic expansion to maximize either precision, recall, or F-measure.

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs for her help in editing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Screenshots explaining how to use the automatic query building tool on the HeTOP website.

[PDF File (Adobe PDF File), 325 KB - medinform_v8i6e12799_app1.pdf]

References

1. National Library of Medicine. Yearly Citation Totals from 2017 MEDLINE/PubMed Baseline: 26,759,399 Citations Found URL: https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_Totals.html [accessed 2018-10-29] [WebCite Cache ID 73Woae0hy]
2. National Library of Medicine. Key MEDLINE® Indicators URL: https://www.nlm.nih.gov/bsd/bsd_key.html [accessed 2018-10-29] [WebCite Cache ID 73WolvSJp]
3. van Dijk N, Hooft L, Wieringa-de Waard M. What are the barriers to residents' practicing evidence-based medicine? A systematic review. Acad Med 2010 Jul;85(7):1163-1170. [doi: 10.1097/ACM.0b013e3181d4152f] [Medline: 20186032]
4. Zwolsman S, te Pas E, Hooft L, Wieringa-de Waard M, van Dijk N. Barriers to GPs' use of evidence-based medicine: a systematic review. Br J Gen Pract 2012 Jul;62(600):e511-e521 [FREE Full text] [doi: 10.3399/bjgp12X652382] [Medline: 22781999]
5. Majid S, Foo S, Luyt B, Zhang X, Theng Y, Chang Y, et al. Adopting evidence-based practice in clinical decision making: nurses' perceptions, knowledge, and barriers. J Med Libr Assoc 2011 Jul;99(3):229-236 [FREE Full text] [doi: 10.3163/1536-5050.99.3.010] [Medline: 21753915]
6. Nelson S. Report on the operation and finance of the Legal Aid Act 1988 for the year. Berlin: HMSO; 2001:978-990.
7. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log. J Am Med Inform Assoc 2007;14(2):212-220 [FREE Full text] [doi: 10.1197/jamia.M2191] [Medline: 17213501]
8. Hoogendam A, Stalenhoef AF, Robbé PF, Overbeke AJ. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. BMC Med Inform Decis Mak 2008 Sep 24;8:42 [FREE Full text] [doi: 10.1186/1472-6947-8-42] [Medline: 18816391]
9. Mosa AS, Yoo I. A study on PubMed search tag usage pattern: association rule mining of a full-day PubMed query log. BMC Med Inform Decis Mak 2013 Jan 09;13:8 [FREE Full text] [doi: 10.1186/1472-6947-13-8] [Medline: 23302604]
10. National Library of Medicine. How PubMed works: Automatic Term Mapping URL: https://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html [accessed 2018-10-29] [WebCite Cache ID 73WpHYH2O]
11. Cooper C, Booth A, Varley-Campbell J, Britten N, Garside R. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Med Res Methodol 2018 Aug 14;18(1):85 [FREE Full text] [doi: 10.1186/s12874-018-0545-3] [Medline: 30107788]
12. Vassar M, Atakpo P, Kash MJ. Manual search approaches used by systematic reviewers in dermatology. J Med Libr Assoc 2016 Oct;104(4):302-304 [FREE Full text] [doi: 10.3163/1536-5050.104.4.009] [Medline: 27822152]
13. Kim S, Fiorini N, Wilbur WJ, Lu Z. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents. J Biomed Inform 2017 Nov;75:122-127 [FREE Full text] [doi: 10.1016/j.jbi.2017.09.014] [Medline: 28986328]
14. PubMed Help. Bethesda, MD: National Center for Biotechnology Information (US); 2005.
15. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 2007 Oct 30;8:423 [FREE Full text] [doi: 10.1186/1471-2105-8-423] [Medline: 17971238]
16. Wei W, Marmor R, Singh S, Demner-Fushman D, Kuo TT, Ohno-Machado L. Finding Related Publicationsxtending the Set of Terms Used to Assess Article Similarity. AMIA Joint Summits on Translational Science Proceedings 2016:234.
17. Afzal M, Hussain M, Ali T, Hussain J, Khan W, Lee S, et al. Knowledge-Based Query Construction Using the CDSS Knowledge Base for Efficient Evidence Retrieval. Sensors (Basel) 2015 Aug 28;15(9):21294-21314 [FREE Full text] [doi: 10.3390/s150921294] [Medline: 26343669]
18. Afzal M, Hussain M, Malik KM, Lee S. Impact of Automatic Query Generation and Quality Recognition Using Deep Learning to Curate Evidence From Biomedical Literature: Empirical Study. JMIR Med Inform 2019 Dec 09;7(4):e13430 [FREE Full text] [doi: 10.2196/13430] [Medline: 31815673]
19. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. Proc AMIA Annu Fall Symp 1997:485-489 [FREE Full text] [Medline: 9357673]
20. Hersh W, Price S, Donohoe L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. Proc AMIA Symp 2000:344-348 [FREE Full text] [Medline: 11079902]
21. Thirion B, Robu I, Darmoni SJ. Optimization of the PubMed Automatic Term Mapping. Stud Health Technol Inform 2009;150:238-242. [Medline: 19745304]

22. Wright T, Ball D, Hersh W. Query expansion using MeSH terms for dataset retrieval: OHSU at the bioCADDIE 2016 dataset retrieval challenge. Database (Oxford) 2017 Jan 01;2017 [FREE Full text] [doi: 10.1093/database/bax065] [Medline: 29220467]
23. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno J, et al. Performance evaluation of Unified Medical Language System®'s synonyms expansion to query PubMed. BMC Med Inform Decis Mak 2012 Feb 29;12:12 [FREE Full text] [doi: 10.1186/1472-6947-12-12] [Medline: 22376010]
24. Xu B, Lin H, Lin Y. Learning to Refine Expansion Terms for Biomedical Information Retrieval Using Semantic Resources. IEEE/ACM Trans Comput Biol and Bioinf 2019 May 1;16(3):954-966. [doi: 10.1109/tcbb.2018.2801303]
25. Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, et al. Health multi-terminology portal: a semantic added-value for patient safety. Stud Health Technol Inform 2011;166:129-138. [Medline: 21685618]
26. Massonnaud C, Lelong R, Kerdelhué G, Lejeune E, Grosjean J, Griffon N, et al. Performance evaluation of three semantic expansions to query PubMed. Health Info Libr J 2019 Dec 14. [doi: 10.1111/hir.12291] [Medline: 31837099]
27. Collins M. NLM Tech Bull. 2019. The New PubMed is Here URL: https://www.nlm.nih.gov/pubs/techbull/nd19/nd19_pubmed_new.html [accessed 2019-05-07]
28. HeTOP, Health Terminology/Ontology Portal. URL: https://www.hetop.eu/hetop/en [accessed 2019-05-07]

Abbreviations

ATM: automatic term mapping
CISMeF: Catalogue et Index des Sites Médicaux de langue Française
CUI: concept unique identifier
HeTOP: Health Terminology/Ontology Portal
KOS: knowledge organization system
MeSH: Medical Subject Headings
NLM: National Library of Medicine
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 12.11.18; peer-reviewed by N Fiorini, W Bramer, A Burgun, M Afzal; comments to author 31.08.19; revised version received 20.01.20; accepted 23.03.20; published 04.06.20.

Please cite as:
Massonnaud CR, Kerdelhué G, Grosjean J, Lelong R, Griffon N, Darmoni SJ
Identification of the Best Semantic Expansion to Query PubMed Through Automatic Performance Assessment of Four Search Strategies on All Medical Subject Heading Descriptors: Comparative Study
JMIR Med Inform 2020;8(6):e12799
URL: https://medinform.jmir.org/2020/6/e12799
doi:10.2196/12799
PMID:32496201

Original Paper

Diabetes Self-Management in the Age of Social Media: Large-Scale Analysis of Peer Interactions Using Semiautomated Methods

Sahiti Myneni[1], PhD; Brittney Lewis[2], MS; Tavleen Singh[1], MS; Kristi Paiva[2], MPH; Seon Min Kim[2], BSc; Adrian V Cebula[2], MS; Gloria Villanueva[2], BSc; Jing Wang[2], PhD, MPH, RN, FAAN

[1]University of Texas School of Biomedical Informatics at Houston, Houston, TX, United States

[2]Center on Smart and Connected Health Technologies, School of Nursing, The University of Texas Health Science Center at San Antonio, San Antonio, TX, United States

Corresponding Author:
Sahiti Myneni, PhD
University of Texas School of Biomedical Informatics at Houston
7000 Fannin Street
Suite 600
Houston, TX, 77030
United States
Phone: 1 7134860115
Email: sahiti.myneni@uth.tmc.edu

Abstract

Background: Online communities have been gaining popularity as support venues for chronic disease management. User engagement, information exposure, and social influence mechanisms can play a significant role in the utility of these platforms.

Objective: In this paper, we characterize peer interactions in an online community for chronic disease management. Our objective is to identify key communications and study their prevalence in online social interactions.

Methods: The American Diabetes Association Online community is an online social network for diabetes self-management. We analyzed 80,481 randomly selected deidentified peer-to-peer messages from 1212 members, posted between June 1, 2012, and May 30, 2019. Our mixed methods approach comprised qualitative coding and automated text analysis to identify, visualize, and analyze content-specific communication patterns underlying diabetes self-management.

Results: Qualitative analysis revealed that “social support” was the most prevalent theme (84.9%), followed by “readiness to change” (18.8%), “teachable moments” (14.7%), “pharmacotherapy” (13.7%), and “progress” (13.3%). The support vector machine classifier resulted in reasonable accuracy with a recall of 0.76 and precision 0.78 and allowed us to extend our thematic codes to the entire data set.

Conclusions: Modeling health-related communication through high throughput methods can enable the identification of specific content related to sustainable chronic disease management, which facilitates targeted health promotion.

(JMIR Med Inform 2020;8(6):e18441) doi:10.2196/18441

KEYWORDS

diabetes; self-management; social media; digital health

Introduction

Background

Diabetes (specifically type 2 diabetes and prediabetes) is a leading public health burden and global health issue. As of 2019, more than 100 million US adults are now living with diabetes or prediabetes [1]. The total estimated cost of diagnosed diabetes in 2020 is $327 billion, including $237 billion in direct medical costs and $90 billion in reduced productivity [1]. Individuals with diagnosed diabetes have annual medical expenditures that are $7900 or approximately 2.3 times higher than they would be in the absence of diabetes ($13,700 vs $5800) [2]. Diabetes can also lead to renal and cardiovascular complications [1]. Addressing lifestyle risk factors, such as poor diet and physical activity, is vital to diabetes prevention and management. Numerous interventions and public health campaigns have been developed to help patients incorporate new behaviors (eg, medication regimen) and modify existing risky behaviors (eg, poor diet) to prevent and manage diabetes (for reviews, see [3-7]). However, the growth rate of diabetes is steady, adding to the health care burden. Adherence to healthy behaviors (eg, proper nutrition) and management of prevailing health conditions (eg, medication adherence) requires a significant support infrastructure that targets individualistic factors and environmental influences for long time intervals [8,9].

Social Relationships and Health Management

Recent research suggests that social relationships play an essential role in an individual's engagement in health issues [10-12]. For example, Christakis and Fowler's analysis of the Framingham data set shows an association between the behavior of members of an individual's social network and the likelihood of smoking cessation [13]. Positive effects of social relationships have been associated with chronic illness self-management [14-17]. Increased levels of social integration are also found to improve the overall wellbeing of individuals [18]. On the other hand, some studies indicate the negative influence of social relationships [19,20]. While community-based social interventions harnessing the positive effects of social contacts exist [21-24], the mechanisms underlying the impact of social relationships on multiple behavioral domains of Diabetes Self-Management (DSM) are not fully understood. Consequently, an understanding of the mechanisms in play for numerous behavioral domains within diabetes management is crucial to promote wellness regimens that can result in sustained adoption.

Online Communities as Secondary Data Sources

The ubiquity of online communities presents us with invaluable data sets in the form of electronic traces of peer interactions [25], which may help to understand social influence in diabetes management. Thanks to the ready availability and accessibility of the internet via mobile phones, peer interactions in online communities often occur in real time. They can provide rich documentation of certain crucial moments in everyday life that influence diabetes prevention and management [26]. Further, it is common for an individual to seek a related online community (eg, newly diagnosed with type 2 diabetes) and navigate the records of peers who have shared their experiences. With the support of online communities and an associated bank of collective knowledge, the individual reflects on the problem, explores available information, and feels able to act, thus eliciting multiple theoretical constructs described in existing models of behavior change ([27-31]. Emerging research shows the complex relationships between online social ties and individuals' self-management of health conditions, thus highlighting the utility of online peer interactions as secondary data sources [17,29]. While we must be cognizant of inferential generalizability [30], these platforms have a tremendous capacity to inform clinicians, behavioral scientists, and technology developers about human health behaviors and ways to harness knowledge from online social media to inform intervention design, content curation, and information dissemination [31-34]. A more in-depth analysis of such interactions provides us with a new lens to inform, enhance, and strengthen existing frameworks of diabetes care delivery, prevention, and management [29,31]. Previous studies on diabetes-related social media interactions have focused on general-purpose platforms such as Twitter and Facebook interventions, where data volume has ranged in the order of hundreds to billions [35-39]. A majority of these studies have attempted to understand the types of diabetes information disseminated, the levels of information spread, and user engagement facilitated by these platforms. However, our understanding of digital environments solely dedicated to diabetes prevention and self-management are quite limited. As such, the semantic context underlying general-purpose and health-specific platforms can vary greatly, consequently affecting the methodological underpinnings of large-scale studies for unpacking the DSM domain in social media.

In this paper, we describe our findings of large-scale analysis of peer interactions in the health-related online community focusing on diabetes management. In addition to abstracting thematic strands underlying peer interactions, we provide a more in-depth analysis of behavior change techniques that manifest in these online discussions using manual coding methods. Further, we extend the reach of qualitative analysis using high throughput computational methods to understand the thematic distribution of peer communication in a diabetes-specific online community. The insights gained from these investigations will enable us to gain a deeper understanding of the digital environment and the nature of the peer interactions they facilitate, inclusive of and beyond social support. Our findings will help us design an enhanced support infrastructure through the development of tailored education interventions and digital solutions that harness social support and influence to promote positive health changes. Such "healthier life" technologies offer considerable advantages over traditional approaches in affordability, scalability, user engagement, and personalization.

Methods

Materials

For this study, we focus on user interactions within the American Diabetes Association (ADA) online community, one of the largest online communities focusing on engaging patients with diabetes and their caregivers in optimizing self-health management [40]. Members are required to have a registered account with the ADA to share content and exchange messages within the online community. The data set spans eight years (2012-2019) and includes publicly available interactions. Behavior before and after diagnosis, treatment effectiveness, healthy behaviors (low carb diet, physical activity), medication adherence, blood glucose self-monitoring, and other topics are discretely captured in this data. For this project, we focused our analysis on type 2–related entries. A total of 80,481 randomly selected de-identified messages exchanged by 1212 members were included in this analysis. We chose type 2 diabetes as the focus of this study because health outcomes and disease management among these patients are impacted by their lifestyle behaviors (diet and physical activity), medication use, and self-monitoring of blood glucose. The research has been reviewed and exempted by the Institutional Review Board at the University of Texas Health Science Center at Houston.

Theme Abstraction

We adopted Directed Content Analysis [41] to identify the core concepts and unifying themes that relate to diabetes prevention and management. First, four independent coders characterized the communication between members of each community, assigning communication themes (inductively derived using grounded theory techniques [42] in our prior work [43]) to randomly selected messages that relate to diabetes prevention and management. Table 1 provides an overview of the qualitative analysis and coding categories. We coded 517 messages to assign thematic labels (shown in Table 1). Each message could have multiple codes applied dependent on the content of the message, and codes were individually and independently assigned by four coders. Each message will have a minimum of two independent coders applying codes. Coders then met and reconciled codes into a master coded document via weekly meeting discussion following iterative comparison and consensus building to ensure objectivity in the coding process. The qualitative analysis allowed us to explain how online platforms are utilized by individual users to mend the gaps in their social and information needs. Also, we conducted a more in-depth analysis of the messages to understand types of social support [44] and the taxonomy of behavior change techniques [45] observed in peer interactions.

Table 1. Sample messages from the American Diabetes Association (ADA) mapped to the communication themes.

Theme	Definition	Sample message snippets from ADA
Social support	Messages where the content reflects the elements of praise, advice, empathy, and guidance	Congratulations on a job well done – and Welcome to the 5% club. Your hard work and persistence paid off. Keep it up. :)
Traditions	Messages that focus on community-specific rituals such as pledges or any engagement practices conducted by moderators or users	How did you do this morning? How've you been doing over time? Nobody knew back then that there would be 28,196 replies to … question. Nobody expected that twice that topic would grow so large that we would have to start over again in a brand new topic to accommodate all those posts.
Teachable moments	Messages that describe incentives to make positive health changes	Stress can have a huge impact on your numbers. Even a single day can raise my numbers significantly and I have had longer periods of stress that I know upped my A1C. So when you are dealing with a stressful time you want to increase your exercise and decrease your carbs.
Obstacles	Messages focusing on hurdles to planned health practices	I did add 3 days of swimming that lasted for 3 months until my swim buddies got on different schedules. I do miss the sun and water so I'm on a search for other swimming holes and buddies. Transportation can be a hurdle, too.
Pharmacotherapy	Messages with explicit discussions on various pharmacotherapy options	Metformin may have a small effect reducing insulin resistance, but its main effect is to keep the liver from sending out too much insulin and over-compensating when blood glucose is a little low, like when it helps to prevent the dawn effect.
Relapse	Messages with descriptions of relapse reasons or confessions	On the issue of my numbers being too high in general… that's a separate issue. I have gotten lax with exercise and eating too many carbs.
Readiness to change	Messages that inspire to initiate positive health changes	I discovered that I had to change "Can't" to "Don't" in my thinking. I "can't" eat that cookie… means "Poor me, someone… is not allowing me to eat that cookie"… I "don't" eat cookies… means that I have a choice it's not something that's part of my life. I am in control.
Cravings	Messages that capture real-time expressions of the urges to deviate from planned health behaviors	Do I miss stuffing my face with pizza or other carbilicious meals? I suppose so, but it's not much of a loss… I miss sugary snacks, I guess that the biggest change.
Alternative medicine	Messages that describe therapies that are not regarded as orthodox by the medical profession	The article has a story of one woman who was getting ready to have a foot/leg amputated (after living with "a terrible wound for 5 years"), but she tried 'the sugar treatment' (my term) and … She ended up not having an amputation.
Progress	Messages in which members communicate their progress based on objective health measures	This summer will mark 8 years since I have been diagnosed with Type 2 diabetes. So far low carb eating, exercise and metformin are keeping me at my target blood glucose numbers.
Patient-reported Outcomes	A message that focuses on subjective progress (positive or negative)	Do I sometimes want to go back? Yes and no. I feel much better now and I know I'm healthier now, so no, I don't want to go back.
Conflict	A message which is argumentative or clarifying a point/topic (not necessarily supportive)	Again I did not say it causes diabetes I said it can cause diabetes – which was the original question. I did not say that there is a direct link between alcoholism and diabetes – but the actions of an alcoholic can contribute to developing diabetes.
Miscellaneous	A message which contains questions or information not about an individual's health status or diabetes management	I'm almost done with my First semester of college. Can you believe that? I did lot hard work.

Automated Methods

Vector representations of all 80,481 messages were generated using distributional semantics methods [46]. The entire data set was then annotated by using the generated vectors as input to a machine learning classifier trained on the manually annotated messages. We exploited recent developments in automated text analysis to measure the extent to which key concepts of interest were expressed within messages between ADA community users, regardless of the specific terms used to express these concepts at the surface level. We applied latent semantic analysis [47], a method of distributional semantics in conjunction with a machine-learning classifier to derive a measure of relatedness between a given message and the previously identified communication themes to estimate the distribution of different types of content across the ADA online community. Ten-fold cross-validation was applied to determine the best performing binary classifier for automating the classification of the entire set of messages. We have used Weka [48] and Semantic Vectors package [49] to build the pipeline for automated classification of ADA peer interactions.

Results

Qualitative Analysis

Theme Abstraction

Based on manual coding of 517 messages, “Social support” was the most common comment theme (n=439, 84.9% of comments), followed by “readiness” (n=97, 18.8%), “teachable moments” (n=76, 14.7%), “pharmacotherapy” (n=71, 13.7%), and “progress” (n=69, 13.3%), “obstacles” (n=48, 9.2% of comments). Additional codes included, “miscellaneous” (n=33, 6.3%), “patient-reported outcomes” (n=29, 5.6%), “traditions” (n=25, 4.8%), “conflict” (n=24, 4.6%), “alternative medicine” (n=7, 1.3%), “relapse” (n=5, 0.97%), and “cravings” (n=1, 0.19%). Given the very nature of the social forum, the majority of the messages exchanged in the ADA community were fostering empathy, affection, and reinforcement that are essential to the sustenance of healthy lifestyle changes. Medication use, motivators for change, and sharing progress also seem to play an important role in diabetes interactions in this community.

Social Support—Anatomical Analysis

A more in-depth analysis of messages specific to social support theme using House taxonomy [39] revealed that the most common form of social support provided was “informational” (n=361, 82.2%), followed by “emotional” (n=155, 35.3%), and “appraisal” (n=9, 0.02%). “Instrumental” support did not apply to our data set, given the lack of manifestation of tangible support (Table 2).

Further analysis revealed the specific behavior change techniques employed by ADA community users. “Social Support,” “Shaping knowledge,” “Feedback and Monitoring,” and “Goals and Planning” were the most utilized behavior change techniques embedded within the messages related to social support theme.

Table 2. Social support analysis.

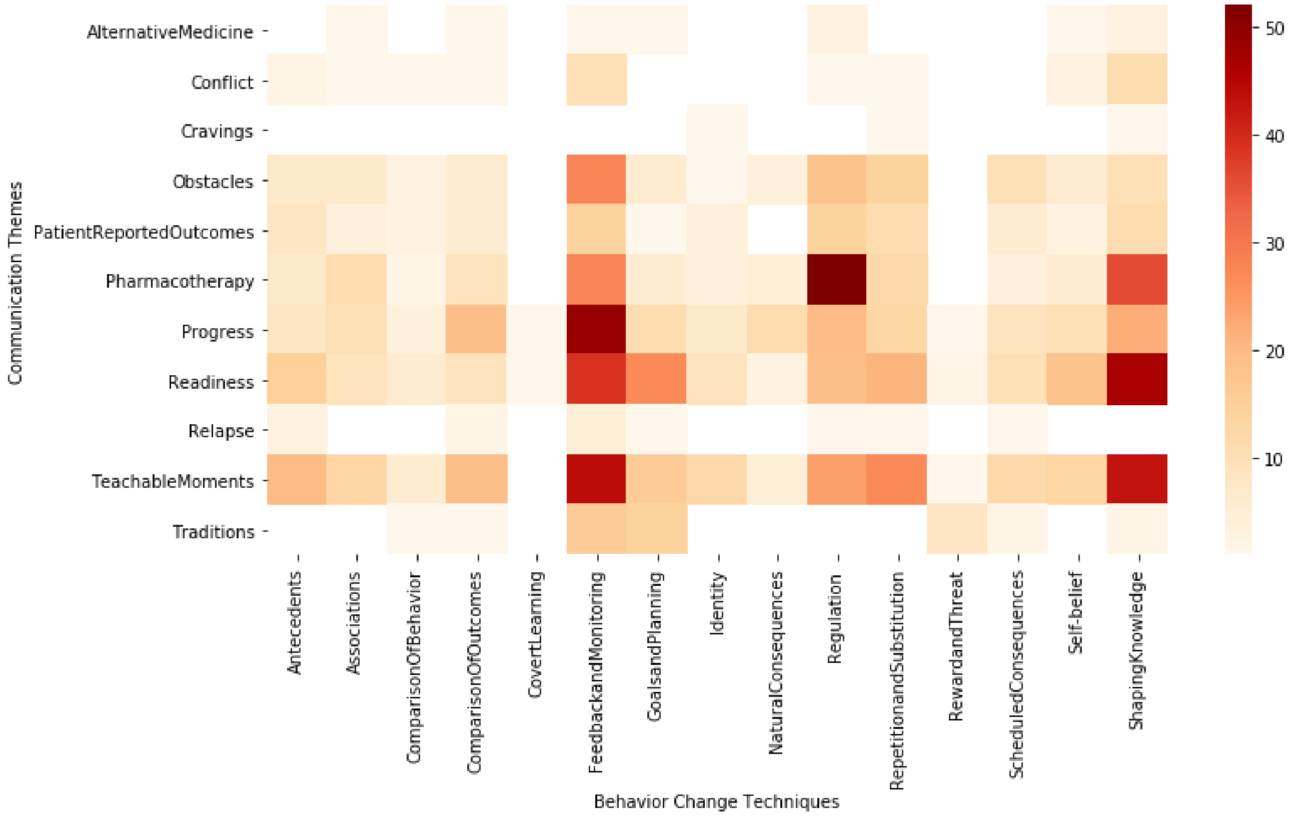
Types of social support	Definition	Example
Informational	Providing advice, suggestions, and information	“I wait for about 6-7 days of bg readings to call a trend for myself when the differences are small, but it is possible over a course of days to note a slight uptick or downtick in bg.”
Emotional	Expressions of empathy, love, trust, and caring	“Way to go …! Congratulations on changing your way of eating and adding in all that exercise.”
Appraisal	Information that is useful for self-evaluation	“Did you ever have diabetes education classes, or consult with a diabetes educator? Do you know how to count carbs? Read here and learn how to make your efforts achieve the best possible outcomes.”
Instrumental	Providing tangible aid and service	N/A[a]

[a] N/A: not applicable.

Beyond Social Support—Anatomical Analysis

Figure 1 shows the thematic dispersion (excluding “Social Support”) across various behavior change techniques, where the color scale represents the number of messages in which a given technique has been observed. “Feedback and monitoring” was the most diversely used technique, followed by “Shaping knowledge,” “Goals and Planning,” and “Repetition and substitution,” and “Regulation.” The least used behavior change techniques include “covert learning,” “rewards and threat,” and “natural consequences.”

Figure 1. Mapping of communication themes and behavior change techniques.



Automated Classification

The precision, recall, and f-measure for the cross-validation of the machine learning technique using the SVM classifier were 0.76, 0.78, and 0.77, respectively. Table 3 provides a summary of the performance for the most commonly used classifiers.

Due to insufficient training examples in the training set, we disregarded 5 of the 13 themes for final classification. Due to a lack of semantic context, we have not included “miscellaneous” in our automated classification system. With the application of our automated classification to the rest of the ADA data set (n=80,481 posts), thematic coverage is as follows: social support (74.2%), readiness to change (12.6%), progress (18.8%), obstacles (10.2%), teachable moments (16.4%), Pharmacotherapy (21.4%), and Patient-reported outcomes (7.1%).

Given the use of high throughput analytical methods to extend manual coding to the rest of the ADA data set, we were able to gain an understanding of the prevalence of DSM-related communication themes in this online community. Understanding thematic prevalence at large scale will now help us with the development of automated support systems using virtual coaching and chatbots for seamless and sustained user experience in online communities such as ADA.

Table 3. Machine learning classifiers applied to peer interactions.

Naïve	Recall	Precision	F-measure
LibLinear	0.64	0.66	0.65
SVM[a]	0.76	0.78	0.77
KNN[b]	0.68	0.65	0.66
J48	0.72	0.66	0.69
Naïve Bayes	0.78	0.54	0.64

[a] SVM: support vector machine.

[b] KNN: k-nearest neighbors.

Discussion

Principal Findings

In this digital era of connected health consumers, the interplay between theory-driven models of diabetes management and observed communication in social media is currently poorly understood [50]. Previous studies have shown that those with DSM who participate in social media forums or platforms saw a decrease in their HbA_{1c} (glycated hemoglobin) [51]. In the future, physicians may “prescribe” a form of social media or platform to reinforce healthy lifestyle choices outside of the clinic.

The results of this study facilitate the ecological analysis of DSM as embedded in peer interactions. This analysis may warrant refining existing models of DSM in the context of face-to-face (rather than online) communication. By using automated social media analysis methods, we will be able to scale up the qualitative analysis to extract relevant communication from large online social media data sets. Though analysis of diabetes management in online health communities is not without precedent [52], prior research does not address methodological scalability and shortcomings to model variances in multiple behaviors and underlying communication attributes in social settings. In this research, we conducted an inductive analysis of DSM strategies, without reliance on a single behavior change theory, as embedded in communication exchanges among members of a health-related online community. This effort enables the extraction of information context significant to behavior change events and social engagement levels in self-management of health-related activities.

Frequent use of online networks for social support, mainly informational, indicates a possible need for individualized diabetes support personnel outside of physician offices. It was noted that users would turn to the online forum to develop a consensus regarding the effectiveness of their medication regimen, exercise routines, and nutritional needs of people with diabetes. A minority of the comments provided solely emotional social support and many comments offered anecdotes to provide context for their diabetes journey. The online forum is a potential method of distributing information regarding their specific illness and sharing new recommendations, as users often share articles and studies they see as relevant or personal experience that helped them better manage their diabetes.

Current research on diabetes prevention and self-management has not addressed the effects of information and social environment. Prior work on content-inclusive network analysis [53-55] provided new methods for modeling network diffusion of communication attributes in online health communities, thus enabling us to disentangle the effects of the theoretical properties of exchanged health information and social structure on health outcomes. With the onset of mobile connectivity in the communication sector, messages exchanged in health-related online communities reflect the intricacies of human health as experienced in real time at the individual, community, and societal levels [33]. The majority of research studies on online health communities focusing on diabetes have analyzed peer-to-peer interactions based on social support categories facilitated by the platforms (eg, informational support, emotional support) [56-58]. However, social support is but one of the numerous interpersonal mechanisms facilitated by the social ties established in online communities. Existing theories of behavior change suggest a myriad of content-driven strategies to elicit specific socio-behavioral mechanisms beyond social support (eg, stimulus control, observational learning) to help individuals change their behavior and self-manage an illness [43,59,60]. Our qualitative analysis of underlying behavior change techniques in peer interactions has highlighted “feedback and monitoring” to be the most used technique, which emphasizes the complex functions of social relationships, which goes beyond the provision of social support. ADA-like platforms can help provide better self-health awareness for individuals through monitoring and knowledge acquisition.

Limitations and Future Work

Our qualitative coding has been limited to inductive analysis and mapping of behavior change techniques in a single online community. Future research should focus on mapping of these inductively derived themes to expansive theory-driven taxonomy such as the Behavior Change Taxonomy [45,60] using computational models for large-scale pattern recognition and identification of independent behavior strands within the DSM in online settings. Further, there may be differences in what is gained from using social media platforms like ADA based on user demographics. Future studies should consider age-specific barriers to information consumption and comprehension in social media platforms. Although we used multiple computational models to perform a large-scale analysis of ADA user interactions, the use of advanced deep learning methods from artificial intelligence research, such as Convolutional Neural Networks [61] and Bidirectional Encoder Representations from Transformers [62], may improve the training of the automated classification system.

Further analysis of peer communication can be deepened through sentiment analysis to find specific emotions in communication, such as anger, happiness, and others. Quantifying sentiments [63] can also help in differentiating their sentiments towards the interventions or other aspects of the behavior change process and regimen. This effort will, in turn, help interventionists identify attitudes and further motivation for user engagement that can arise from satisfaction/dissatisfaction with the intervention.

Conclusions

Behavior modification, such as balanced nutrition, an increase in physical activity, and medication adherence, is a critical component of DSM. Patient engagement in DSM consists of the adoption of healthy behaviors and abstinence from risky behaviors. However, the modification of such behaviors is challenging. Numerous public health efforts have been made to promote healthy behaviors over the years, but their utility and efficacy have been suboptimal. The utility of online social media to foster behavior change has been recognized as one sustainable solution. However, little is known about how we can harness social platforms to facilitate positive changes and promote DSM. Health-related online communities present a unique opportunity to improve our understanding of such socio-behavioral mechanisms, as communication in this context is digitally archived, permitting analysis of the dynamics of social influence as they manifest in peer interactions. Our methods have allowed us to abstract the essence of peer-to-peer communication in online communities at scale and to elucidate ways in which observable digital interactions relate to behavior modification endeavors as related to diabetes prevention and management. Our findings will provide the basis for an integrated approach to the problem of chronic disease management and underlying subtasks of behavior change. Such work will have implications for the design of behavior support technologies that offer automated personalization to improve self-management behaviors at the individual and population level.

Acknowledgments

The research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Numbers 1R01LM012974-01A1, and University of Texas Health San Antonio Center on Smart and Connected Health Technologies pilot funding. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

None declared.

References

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Estimates of Diabetes and Its Burden in the United States. Atlanta, GA. URL: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.Pdf [accessed 2020-05-30]
2. Herman WH. The economic costs of diabetes: is it time for a new treatment paradigm? Diabetes Care 2013 Apr 21;36(4):775-776 [FREE Full text] [doi: 10.2337/dc13-0270] [Medline: 23520368]
3. Norris SL, Engelgau MM, Narayan KM. Effectiveness of self-management training in type 2 diabetes: a systematic review of randomized controlled trials. Diabetes Care 2001 Mar 01;24(3):561-587. [doi: 10.2337/diacare.24.3.561] [Medline: 11289485]
4. Gillies CL, Abrams KR, Lambert PC, Cooper NJ, Sutton AJ, Hsu RT, et al. Pharmacological and lifestyle interventions to prevent or delay type 2 diabetes in people with impaired glucose tolerance: systematic review and meta-analysis. BMJ 2007 Feb 10;334(7588):299 [FREE Full text] [doi: 10.1136/bmj.39063.689375.55] [Medline: 17237299]
5. Renders CM, Valk GD, Griffin SJ, Wagner EH, Eijk Van JT, Assendelft WJ. Interventions to improve the management of diabetes in primary care, outpatient, and community settings: a systematic review. Diabetes Care 2001 Oct 01;24(10):1821-1833. [doi: 10.2337/diacare.24.10.1821] [Medline: 11574449]
6. Willey C, Redding C, Stafford J, Garfield F, Geletko S, Flanigan T, et al. Stages of change for adherence with medication regimens for chronic disease: Development and validation of a measure. Clinical Therapeutics 2000 Jul;22(7):858-871. [doi: 10.1016/s0149-2918(00)80058-2]
7. Strecher V, Wang C, Derry H, Wildenhaus K, Johnson C. Tailored interventions for multiple risk behaviors. Health education research Oct 1 2002;17(5):619-626. [doi: 10.1093/her/17.5.619]
8. Gallant MP. The influence of social support on chronic illness self-management: a review and directions for research. Health Educ Behav 2003 Apr;30(2):170-195. [doi: 10.1177/1090198102251030] [Medline: 12693522]
9. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Social Science & Medicine 2000 Sep;51(6):843-857. [doi: 10.1016/s0277-9536(00)00065-4]
10. Blanchard CG, Albrecht TL, Ruckdeschel JC, Grant CH, Hemmick RM. The Role of Social Support in Adaptation to Cancer and to Survival. Journal of Psychosocial Oncology 1995 Aug 15;13(1-2):75-95. [doi: 10.1300/j077v13n01_05]
11. Burg MM, Seeman TE. Families and Health: the Negative Side of Social Ties. Annals of Behavioral Medicine 1994;16(2):109-115.
12. Christakis NA, Fowler JH. The Collective Dynamics of Smoking in a Large Social Network. N Engl J Med 2008 May 22;358(21):2249-2258. [doi: 10.1056/nejmsa0706154]
13. Tillotson LM, Smith MS. Locus of control, social support, and adherence to the diabetes regimen. Diabetes Educ 1996 Jun 30;22(2):133-139. [doi: 10.1177/014572179602200206] [Medline: 8697963]
14. Nam S, Chesla C, Stotts NA, Kroon L, Janson SL. Barriers to diabetes management: patient and provider factors. Diabetes Res Clin Pract 2011 Jul;93(1):1-9. [doi: 10.1016/j.diabres.2011.02.002] [Medline: 21382643]
15. Lo R. Correlates of expected success at adherence to health regimen of people with IDDM. J Adv Nurs 1999 Aug 25;30(2):418-424. [doi: 10.1046/j.1365-2648.1999.01085.x] [Medline: 10457244]
16. Maclean HM. Patterns of diet related self-care in diabetes. Social Science & Medicine 1991 Jan;32(6):689-696. [doi: 10.1016/0277-9536(91)90148-6]
17. Poirier J, Cobb NK. Social influence as a driver of engagement in a web-based health intervention. J Med Internet Res 2012 Feb 22;14(1):e36 [FREE Full text] [doi: 10.2196/jmir.1957] [Medline: 22356829]
18. Kaplan RM, Toshima MT. The Functional Effects of Social Relationships on Chronic Illnesses and Disability. In: Sarason BR, Sarason IG, Pierce GR, editors. Wiley series on personality processes. Social support: An interactional view. Hoboken, NJ: John Wiley & Sons; 1990:427-453.
19. Wortman CB, Conway TL. The role of social support in adaptation and recovery from physical illness. In: Cohen S, Syme SL, editors. Social Support and Health. Orlando, FL: Academic Press; 1985:281-302.

20. Alexander C, Piazza M, Mekos D, Valente T. Peers, schools, and adolescent cigarette smoking. Journal of Adolescent Health 2001 Jul;29(1):22-30. [doi: 10.1016/s1054-139x(01)00210-5]
21. van der Eijk M, Faber MJ, Aarts JW, Kremer JA, Munneke M, Bloem BR. Using online health communities to deliver patient-centered care to people with chronic conditions. J Med Internet Res 2013 Jun 25;15(6):e115 [FREE Full text] [doi: 10.2196/jmir.2476] [Medline: 23803284]
22. Myneni S, Fujimoto K, Cohen T. Leveraging Social Media for Health Promotion and Behavior Change: Methods of Analysis and Opportunities for Intervention. In: Cognitive Informatics in Health and Biomedicine. Cham: Springer; 2017:315-345.
23. Maloney-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Trans. Comput.-Hum. Interact 2005 Jun;12(2):201-232. [doi: 10.1145/1067860.1067864]
24. Ferguson T. Health online: How to find health information, support groups, and self-help communities in cyberspace. Reading, MA: Addison-Wesley; 1996.
25. Eysenbach G, Powell J, Englesakis M, Rizo C, Stern A. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. BMJ 2004 May 15;328(7449):1166 [FREE Full text] [doi: 10.1136/bmj.328.7449.1166] [Medline: 15142921]
26. Heron K, Smyth J. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. British journal of health psychology 2010;15(1):1-39. [doi: 10.1348/135910709x466063]
27. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing health data for better outcomes on PatientsLikeMe. J Med Internet Res 2010 Jun 14;12(2):e19 [FREE Full text] [doi: 10.2196/jmir.1549] [Medline: 20542858]
28. Lefebvre RC, Bornkessel AS. Digital Social Networks and Health. Circulation 2013 Apr 30;127(17):1829-1836. [doi: 10.1161/circulationaha.112.000897]
29. Centola D. Social Media and the Science of Health Behavior. Circulation 2013 May 28;127(21):2135-2144. [doi: 10.1161/circulationaha.112.101816]
30. Tufekci Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. ICWSM 2014:505-514.
31. Cobb NK, Graham AL. Health behavior interventions in the age of facebook. Am J Prev Med 2012 Nov;43(5):571-572. [doi: 10.1016/j.amepre.2012.08.001] [Medline: 23079184]
32. Graham A, Cobb C, Cobb N. The internet, social media, health decision-making. In: Handbook of Health Decision Science. New York, NY: Springer; 2016:335-355.
33. van Mierlo T, Li X, Hyatt D, Ching AT. Demographic and Indication-Specific Characteristics Have Limited Association With Social Network Engagement: Evidence From 24,954 Members of Four Health Care Support Groups. J Med Internet Res 2017 Feb 17;19(2):e40 [FREE Full text] [doi: 10.2196/jmir.6330] [Medline: 28213340]
34. Jane M, Hagger M, Foster J, Ho S, Pal S. Social media for health promotion and weight management: a critical debate. BMC Public Health 2018 Jul 28;18(1):932 [FREE Full text] [doi: 10.1186/s12889-018-5837-3] [Medline: 30055592]
35. AlQarni ZA, Yunus F, Househ MS. Health information sharing on Facebook: An exploratory study on diabetes mellitus. J Infect Public Health 2016 Nov;9(6):708-712 [FREE Full text] [doi: 10.1016/j.jiph.2016.08.015] [Medline: 27618634]
36. Backa KE, Holmberg K, Ek S. Communicating diabetes and diets on Twitter - a semantic content analysis. IJNVO 2016;16(1):8. [doi: 10.1504/ijnvo.2016.075133]
37. Karami A, Dahl AA, Turner-McGrievy G, Kharrazi H, Shaw G. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. International Journal of Information Management 2018 Feb;38(1):1-6. [doi: 10.1016/j.ijinfomgt.2017.08.002]
38. Rus HM, Cameron LD. Health Communication in Social Media: Message Features Predicting User Engagement on Diabetes-Related Facebook Pages. Ann Behav Med 2016 Oct 8;50(5):678-689. [doi: 10.1007/s12160-016-9793-9] [Medline: 27059761]
39. Liu Y, Mei Q, Hanauer DA, Zheng K, Lee JM. Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-Related Tweets. JMIR Diabetes 2016 Nov 07;1(2):e4 [FREE Full text] [doi: 10.2196/diabetes.6256] [Medline: 30291053]
40. American Diabetes Association. URL: https://community.diabetes.org/home [accessed 2020-05-30]
41. 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]
42. Mattley C, Strauss A, Corbin J. Grounded Theory in Practice. Contemporary Sociology 1999 Jul;28(4):489. [doi: 10.2307/2655359]
43. Myneni S, Cobb N, Cohen T. In Pursuit of Theoretical Ground in Behavior Change Support Systems: Analysis of Peer-to-Peer Communication in a Health-Related Online Community. J Med Internet Res 2016 Feb 02;18(2):e28 [FREE Full text] [doi: 10.2196/jmir.4671] [Medline: 26839162]
44. House JS. Work, stress, and social support. In: Addison-Wesley series on occupational stress. Boston, MA: Addison-Wesley; 1981.
45. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug 20;46(1):81-95. [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]

46. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. J Biomed Inform 2009 Apr;42(2):390-405 [FREE Full text] [doi: 10.1016/j.jbi.2009.02.002] [Medline: 19232399]
47. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes 1998 Jan;25(2-3):259-284. [doi: 10.1080/01638539809545028]
48. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. SIGKDD Explor. Newsl 2009 Nov 16;11(1):10-18. [doi: 10.1145/1656274.1656278]
49. Widdows D, Ferraro K. Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Paris: European Language Resources Association (ELRA); May 2008.
50. Riley WT, Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R. Health behavior models in the age of mobile interventions: are our theories up to the task? Transl Behav Med 2011 Mar 24;1(1):53-71 [FREE Full text] [doi: 10.1007/s13142-011-0021-7] [Medline: 21796270]
51. Alcántara-Aragón V. Improving patient self-care using diabetes technologies. Ther Adv Endocrinol Metab 2019;10:2042018818824215 [FREE Full text] [doi: 10.1177/2042018818824215] [Medline: 30728941]
52. Greene JA, Choudhry NK, Kilabuk E, Shrank WH. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen Intern Med 2011 Mar 13;26(3):287-292 [FREE Full text] [doi: 10.1007/s11606-010-1526-3] [Medline: 20945113]
53. Myneni S, Cobb N, Cohen T. Finding meaning in social media: content-based social network analysis of QuitNet to identify new opportunities for health promotion. Studies in health informatics and technology 2013:807-811.
54. Myneni S, Cobb NK, Cohen T. Content-specific network analysis of peer-to-peer communication in an online community for smoking cessation. AMIA Annu Symp Proc 2016;2016:934-943 [FREE Full text] [Medline: 28269890]
55. Myneni S, Fujimoto K, Cobb N, Cohen T. Content-Driven Analysis of an Online Community for Smoking Cessation: Integration of Qualitative Techniques, Automated Text Analysis, and Affiliation Networks. Am J Public Health 2015 Jun;105(6):1206-1212. [doi: 10.2105/ajph.2014.302464]
56. Johnson C, Feinglos M, Pereira K, Hassell N, Blascovich J, Nicollerat J, et al. Feasibility and preliminary effects of a virtual environment for adults with type 2 diabetes: pilot study. JMIR Res Protoc 2014 Apr 08;3(2):e23 [FREE Full text] [doi: 10.2196/resprot.3045] [Medline: 24713420]
57. Shaw RJ, Johnson CM. Health Information Seeking and Social Media Use on the Internet among People with Diabetes. Online J Public Health Inform 2011 Jun 22;3(1) [FREE Full text] [doi: 10.5210/ojphi.v3i1.3561] [Medline: 23569602]
58. Hilliard M, Sparling K, Hitchcock J, Oser T, Hood K. The emerging diabetes online community. Curr Diabetes Rev 2015 Jul 29;11(4):261-272 [FREE Full text] [doi: 10.2174/1573399811666150421123448] [Medline: 25901500]
59. Glanz K, Rimer B, Viswanath K. Health Behavior and Health Education: Theory, Research, and Practice. Hoboken, NJ: John Wiley & Sons; 2008.
60. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
61. Kim Y. Convolutional neural networks for sentence classification. In: In Proceedings of EMNLP. 2014 Presented at: EMNLP; October 25-29; Doha, Qatar p. 25.
62. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2019 Presented at: Association for Computational Linguistics; June 2-7; Minneapolis, Minnesota p. 4171-4186.
63. Crossley SA, Kyle K, McNamara DS. Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behav Res 2016 May 18;49(3):803-821. [doi: 10.3758/s13428-016-0743-z]

Abbreviations

ADA: American Diabetes Association
DSM: diabetes self-management
HbA_{1c}: glycated hemoglobin
SVM: support vector machine
KNN: k-nearest neighbors

Edited by G Eysenbach; submitted 26.02.20; peer-reviewed by K Radhakrishnan, H Stuckey, E Da Silva, K Pretorius; comments to author 18.03.20; revised version received 14.05.20; accepted 04.06.20; published 30.06.20.

Please cite as:
Myneni S, Lewis B, Singh T, Paiva K, Kim SM, Cebula AV, Villanueva G, Wang J
Diabetes Self-Management in the Age of Social Media: Large-Scale Analysis of Peer Interactions Using Semiautomated Methods
JMIR Med Inform 2020;8(6):e18441
URL: https://medinform.jmir.org/2020/6/e18441
doi:10.2196/18441
PMID:32602843

Original Paper

Evaluating the Representativeness of US Centricity Electronic Medical Records With Reports From the Centers for Disease Control and Prevention: Comparative Study on Office Visits and Cardiometabolic Conditions

Olga Montvida[1], PhD; John Epoh Dibato[1], MSc; Sanjoy Paul[1], PhD

Melbourne EpiCentre, University of Melbourne, Melbourne, Australia

Corresponding Author:
Sanjoy Paul, PhD
Melbourne EpiCentre
University of Melbourne
The Royal Melbourne Hospital – City Campus
7 East, Main Building Grattan Street, Parkville Victoria
Melbourne, 3050
Australia
Phone: 61 0435659875
Fax: 61 393428780
Email: sambhupaul@hotmail.com

Abstract

Background: Electronic medical record (EMR)–based clinical and epidemiological research has dramatically increased over the last decade, although establishing the generalizability of such big databases for conducting epidemiological studies has been an ongoing challenge. To draw meaningful inferences from such studies, it is essential to fully understand the characteristics of the underlying population and potential biases in EMRs.

Objective: This study aimed to assess the generalizability and representativity of the widely used US Centricity Electronic Medical Record (CEMR), a primary and ambulatory care EMR for population health research, using data from the National Ambulatory Medical Care Surveys (NAMCS) and the National Health and Nutrition Examination Surveys (NHANES).

Methods: The number of office visits reported in the NAMCS, designed to meet the need for objective and reliable information about the provision and the use of ambulatory medical care services, was compared with similar data from the CEMR. The distribution of major cardiometabolic diseases in the NHANES, designed to assess the health and nutritional status of adults and children in the United States, was compared with similar data from the CEMR.

Results: Gender and ethnicity distributions were similar between the NAMCS and the CEMR. Younger patients (aged <15 years) were underrepresented in the CEMR compared with the NAMCS. The number of office visits per 100 persons per year was similar: 277.9 (95% CI 259.3-296.5) in the NAMCS and 284.6 (95% CI 284.4-284.7) in the CEMR. However, the number of visits for males was significantly higher in the CEMR (CEMR: 270.8 and NAMCS: 239.0). West and South regions were underrepresented and overrepresented, respectively, in the CEMR. The overall prevalence of diabetes along with age and gender distribution was similar in the CEMR and the NHANES: overall prevalence, 10.1% and 9.7%; male, 11.5% and 10.8%; female, 9.1% and 8.8%; age 20 to 40 years, 2.5% and 1.8%; and age 40 to 60 years, 9.4% and 11.1%, respectively. The prevalence of obesity was similar: 42.1% and 39.6%, with similar age and female distribution (41.5% and 41.1%) but different male distribution (42.7% and 37.9%). The overall prevalence of high cholesterol along with age and female distribution was similar in the CEMR and the NHANES: overall prevalence, 12.4% and 12.4%; and female, 14.8% and 13.2%, respectively. The overall prevalence of hypertension was significantly higher in the CEMR (33.5%) than in the NHANES (95% CI: 27.0%-31.0%).

Conclusions: The distribution of major cardiometabolic diseases in the CEMR is comparable with the national survey results. The CEMR represents the general US population well in terms of office visits and major chronic conditions, whereas the potential subgroup differences in terms of age and gender distribution and prevalence may differ and, therefore, should be carefully taken care of in future studies.

(JMIR Med Inform 2020;8(6):e17174) doi:10.2196/17174

KEYWORDS

electronic medical records; observational study; epidemiology; population health

Introduction

Background

Large national surveys and registry data provide epidemiological and population-level health information. Although such studies will remain as gold standards in evaluating the health state at a population level, the more recent development of large real-world data (RWD) from electronic medical records (EMRs) and claims data for therapeutic management and population-level safety evaluations provide additional and unique opportunities to expand our understanding in a broad class of clinical, epidemiological, and public health–related questions [1-6].

EMR data are collected during routine medical care, offering the opportunity to investigate clinical questions from a real-world perspective. Although randomized clinical trials (RCTs) allow the evaluation of the safety and efficacy of interventions in a design-led population, the EMR-based studies allow for comparative effectiveness and safety studies, apart from revolutionizing the approach to efficient pharmacovigilance. RWD-based studies also provide opportunities to explore clinical questions in populations that are often excluded from RCTs, such as pregnant, older, or comorbid patients. Furthermore, real-world studies allow us to investigate questions that may be unethical for testing in RCTs. EMRs are also used to track how clinical guidelines are implemented in real-world practices and to research the quality of clinical care.

The epidemiological value of EMR-based research directly depends on the size of the EMR network. Several EMR systems were implemented at the national level, and most familiar representatives include databases from the United States, the United Kingdom, Sweden, Norway, and Denmark [7-10]. The representativeness of some of these databases in terms of demographics and chronic and rare diseases has been shown in some studies [8,9,11-14].

Apart from health research based on data from individual practices, pharmacies, insurers, claims, or prescriptions, the MarketScan Commercial Claims and Encounters Database, owned by Truven Health Analytics, is one of the most commonly used data source for health research in the United States [15,16]. The Veteran Affairs–integrated health care system is another widely used data source in the United States [17,18]. One of the oldest primary and ambulatory EMR systems in the United States is the Centricity Electronic Medical Record (CEMR), owned by General Electric, which provides an opportunity for research using deidentified data on more than 45 million patients from all states of the United States [19,20].

The CEMR database has been extensively used for health outcome academic research worldwide in the fields of diabetes [21-26], cardiovascular research [27-31], obesity [32-34], inflammatory diseases [35-38], mental health [39-41], and other diseases [42,43]. To draw meaningful inferences from such studies and to generalize the results, it is essential to understand the underlying population. For instance, using the CEMR, Montvida et al [44] described trends in the antidiabetic drug prescription patterns during the years 2005-2016. However, the study should be interpreted in the light of overall representativity of the CEMR with respect to diabetes as well as gender and ethnic differences, as all these factors affect drug choices.

To the best of our knowledge, only two studies have investigated the representativeness of the CEMR database [14,45]. Brixner et al [14] evaluated the BMI and laboratory data from the CEMR in 2003 to 2004 in comparison with the US national health surveys. Crawford et al [45] compared the National Ambulatory Medical Care Survey (NAMCS) with CEMR's office visits during 2005 and concluded that CEMR data provide a more accurate estimate of the distribution of diagnoses in ambulatory visits in the United States.

Aims

Given the significant increase in CEMR coverage since the last report was published and exponentially increasing volume of RWD-based research, we aimed to repeat and expand the exploration of the generalizability and representativity of the CEMR database with two of the most widely used and relevant survey results from the United States. Specifically, the goals of this study were to compare (1) patient demographics in the CEMR with the NAMCS and (2) the prevalence of obesity, hypertension, high total cholesterol, and diabetes in the CEMR with the respective reports based on the National Health and Nutrition Examination Surveys (NHANES).

Data

Centricity Electronic Medical Record

The CEMR incorporates patient-level data from independent physician practices, academic medical centers, hospitals, and large integrated delivery networks in the United States. The Medical Quality Improvement Consortium is a rapidly growing community that contributes deidentified clinical data to the CEMR research database to enable quality improvement, benchmarking, and population-based medical research [40,46]. With an average follow-up of 4.5 years, the CEMR research database covers more than 35,000 health care providers from all states of the United States, where approximately 70% are primary care providers. Longitudinal EMRs were available for more than 45 million individuals from 1995 to September 2018, with comprehensive patient-level information on demographics, anthropometric measures, disease events, medications, and clinical and laboratory measures. The database has been extensively used in academic research [14,44,45].

The National Ambulatory Medical Care Survey

A report based on the 2016 NAMCS data was used in this study [47]. The excerpts from the Centers for Disease Control and

Prevention (CDC) website are presented in the following two paragraphs [47,48].

The NAMCS is a national survey designed to meet the need for objective, reliable information about the provision and the use of ambulatory medical care services in the United States. The findings were based on a sample of visits to nonfederally employed office-based physicians who are primarily engaged in direct patient care. Physicians specializing in anesthesiology, pathology, and radiology were excluded from the survey. Each physician was randomly assigned to a 1-week reporting period. During this period, data for a systematic random sample of visits were recorded by Census interviewers using an automated patient record form (PRF) developed for that purpose.

The 2016 NAMCS sampling design used a stratified two-stage sample, with physicians selected in the first stage and visits in the second stage. The 2016 NAMCS sample included 3699 physicians. Of the 2080 in-scope (eligible) physicians, 677 completed PRFs in the study. Of the 677 physicians who completed PRFs, 536 participated fully or adequately (ie, at least one half of the expected PRFs were submitted, based on the total number of visits during the reporting week) and 141 participated minimally (ie, fewer than half of the expected number of PRFs were submitted). Within physician practices, data were abstracted from medical records for up to 30 sampled visits during a randomly assigned 1-week reporting period. In total, 13,165 PRFs were submitted. The participation rate—the percentage of in-scope physicians for whom at least one PRF was completed—was 39.3%. The response rate—the percentage of in-scope physicians for whom at least one half of their expected number of PRFs was completed—was 32.7%. Among the 4 census regions, response rates ranged from 24.6% to 40.0%.

The National Health and Nutrition Examination Survey

The NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States [49]. The survey consists of interviews conducted in participants' homes, standardized physical examinations in mobile examination centers, and laboratory tests on blood and other specimens.

Individual reports produced by the CDC based on the NHANES 2015-2016 data were used to compare the prevalence of obesity [50], hypertension [51], and high total cholesterol [52]. The latest CDC report for diabetes prevalence was based on the NHANES 2013-2016 data [53].

Methods

Data from the CEMR were matched on methods to individual CDC reports as close as possible.

Office Visits

Data on percent distribution of office visits and number of office visits per 100 patients by various subgroups from the NAMCS were compared with similar data from the CEMR. All office visits in 2016 for patients with nonmissing age and sex from the CEMR were aggregated to match the NAMCS report as close as possible.

Obesity

In the NHANES, obesity was defined as a BMI of 30 kg/m^2 or greater for adults aged 20 years and older [50].

In the CEMR, the proportion of obese people was estimated among people aged older than 20 years and with at least one BMI measure (direct or estimated using weight and height) during the years 2015-2016. Women who had pregnancy-related records before or within the estimated time frame were excluded.

Hypertension

In the NHANES, systolic blood pressure (SBP) of 140 mm Hg or greater, diastolic blood pressure (DBP) of 90 mm Hg or greater, or currently taking medication to lower high blood pressure were defining hypertension for people aged 18 years and older [51].

In the CEMR, the proportion of patients with hypertension during the years 2015-2016 was estimated among people aged older than 18 years. On average, patients had 4 blood pressure measures during a 2-year time frame. Those who had an average of available measures for SBP of 140 mm Hg or greater, those who had an average of available measures for DBP of 90 mm Hg or greater, or those who were taking medication to lower high blood pressure during the respective time frame were considered to have hypertension. Blood pressure–lowering medications included diuretics, peripheral vasodilators, beta blockers, calcium channel blockers, angiotensin-converting enzyme inhibitors, angiotensin II receptor blockers, and other agents acting on the renin-angiotensin system. Only medications that are indicated to lower blood pressure were preserved within these drug classes.

High Total Cholesterol

In the NHANES, proportions of participants aged 20 years and older with high total cholesterol (≥240 mg/dL) were reported [52]. In the CEMR, among people aged older than 20 years and with at least one available cholesterol measure, the proportions of those with total cholesterol of 240 mg/dL or greater were estimated during the years 2015-2016.

Diabetes

In the NHANES, participants were classified as having diagnosed diabetes if they answered "yes" to the question, "Other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?" [53]. Participants were classified as having undiagnosed diabetes if they did not report a diagnosis of diabetes by a health care provider and their fasting (8-24 hours) plasma glucose level was 126 mg/dL or greater or their hemoglobin A_{1c} (HbA_{1c}) level was 6.5% or greater. Participants were randomly assigned to a morning, afternoon, or evening examination. Fasting plasma glucose data from the morning examination (after an 8- to 24-hour fast) were used to define total and undiagnosed diabetes.

In the CEMR, an algorithm to identify patients with diabetes was developed on the basis of (1) diabetes diagnostic codes (International Classification of Diseases and SNOMED), (2) antidiabetic medication prescription patterns, (3) availability of

2 measurements of HbA_{1c} level of 6.5% or greater or fasting blood glucose level of 126 mg/dL or greater or random blood glucose level of 200 mg/dL or greater within 1 year, and (4) keyword searching procedures for diabetic-related terms from the clinical notes of every patient. The algorithm was developed on the basis of clinical guidelines and machine learning suggestions described by Adjah et al [54] for a database from the United Kingdom. Patients who were prescribed metformin for polycystic ovary syndrome were detected and excluded. In the case of nondefinite diabetes subtype, a patient's age and insulin and noninsulin prescription patterns were used to distinguish subtypes. The off-label use of antidiabetic drugs was not explored. For analyses in this study, patients with prediabetes and gestational diabetes were excluded. The proportion of patients with coded and noncoded diabetes was estimated among adults aged 20 years and older and who were active in the CEMR during the years 2013-2016.

Statistical Methods

Proportional distributions between the CEMR and the NAMCS and NHANES were compared using the chi-square test, where appropriate. Office visit estimates in the NAMCS report are based on sample data weighted to produce annual national estimates and include SEs. All estimates in the NHANES reports were age adjusted using the 2000 US Census population. For the NAMCS and the NHANES estimates, 95% CIs were calculated using the available SE estimates from the reports. Crude estimates from the CEMR data were calculated and presented in this study, 95% CI for percentages were calculated based on binomial distribution assumption, and 95% CIs for number of visits per 100 persons per year were calculated assuming a Poisson distribution.

Statistical equivalence for the pairwise comparisons of proportional distributions, where appropriate, was evaluated using the two one-sided test (TOST) of equivalence [55,56] with a ±2.5, ±5, and ±7.5 percentage point equivalence margins. Using population summary statistics on mean, SD, and total number of office visits, the TOST procedure available in SAS 9.4 (SAS Institute Inc) was employed.

Results

Office Visits

The NAMCS estimated 883,725,000 office visits in the United States in 2016. In the CEMR, 29,207,860 office visits in 2016 occurred for patients with nonmissing age and sex. In the NAMCS and the CEMR, sex distribution was similar (equivalent at 2.5 percentage point margin), where 58.0% (95% CI 56.2-59.8) and 59.8% (95% CI 59.8-59.8) of all visits were by females (Table 1). The number of visits per 100 females per year was similar in the NAMCS and the CEMR: 315.0 (95% CI 291.5-338.5) versus 294.6 (95% CI 294.5-294.8), whereas the number of visits per 100 males per year was lower in the NAMCS compared with the CEMR: 239.0 (220.6-257.4) versus 270.8 (270.6-270.9).

Table 1. Patient characteristics at office visits (National Ambulatory Medical Care Surveys estimated visits, N=883,725,000 and Centricity Electronic Medical Record total visits, N=29,207,860).

Characteristics[a]	National Ambulatory Medical Care Surveys 2016		Centricity Electronic Medical Record 2016	
	Percent distribution (95% CI)	Number of visits per 100 persons per year (95% CI)	Percent distribution (95% CI)	Number of visits per 100 persons per year (95% CI)
All visits	100 (N/A)[b]	277.9 (259.3-296.5)	100 (N/A)	284.6 (284.4-284.7)
Age (years)				
<15	17.7 (14.6-20.8)	257.4 (205.5-309.3)	11.2 (11.2-11.2)[c]	276.5 (276.2-276.8)
15-24	7.4 (6.6-8.2)	153.0 (132.6-173.4)	7.5 (7.4-7.5)	244.8 (244.5-245.1)
25-44	19.3 (17.5-21.1)	205.4 (184.4-226.4)	20.0 (20.0-20.0)	272.1 (271.9-272.4)
45-64	28.5 (26.7-30.3)	301.9 (275.0-328.8)	30.9 (30.9-31.0)	280.3 (280.1-280.4)
65-74	15.0 (13.8-16.2)	465.2 (419.7-510.7)	16.3 (16.3-16.3)	303.3 (303.1-303.6)
≥75	12.1 (10.9-13.3)	546.8 (491.3-602.3)	14.1 (14.1-14.1)	329.1 (328.8-329.4)
Female				
Total	58.0 (56.2-59.8)	315.0 (291.5-338.5)	59.8 (59.8-59.8)	294.6 (294.5-294.8)
<15	8.4 (6.6-10.2)	247.7 (186.9-308.5)	5.4 (5.3-5.4)[d]	274.4 (274.0-274.8)
15-24	4.6 (3.8-5.4)	194.4 (160.9-227.9)	4.9 (4.9-4.9)	267.3 (266.8-267.7)
25-44	13.4 (12.0-14.8)	281.8 (249.1-314.5)	13.8 (13.8-13.8)	293.8 (293.5-294.0)
45-64	16.4 (15.0-17.8)	337.2 (303.1-371.3)	18.2 (18.2-18.2)	286.5 (286.3-286.8)
65-74	8.0 (7.2-8.8)	467.9 (415.6-520.2)	9.2 (9.2-9.2)	310.0 (309.6-310.4)
≥75	7.1 (6.3-7.9)	549.9 (484.2-615.6)	8.3 (8.2-8.3)	335.2 (334.8-335.6)
Male				
Total	42.0 (40.2-43.8)	239.0 (220.6-257.4)	40.2 (40.2-40.2)	270.8 (270.6-270.9)
<15	9.4 (7.8-11.0)	266.7 (216.3-317.1)	5.9 (5.8-5.9)[d]	278.5 (278.1-278.9)
15-24	2.7 (2.3-3.1)	112.3 (93.9-130.7)	2.5 (2.5-2.5)	210.0 (209.5-210.5)
25-44	5.9 (5.1-6.7)	126.9 (108.7-145.1)	6.2 (6.2-6.2)	233.6 (233.3-233.9)
45-64	12.1 (10.9-13.3)	264.4 (234.4-294.4)	12.7 (12.7-12.7)	271.7 (271.5-272.0)
65-74	6.9 (6.1-7.7)	462.2 (410.1-514.3)	7.1 (7.1-7.1)	295.1 (294.7-295.5)
≥75	5 (4.4-5.6)	542.4 (479.5-605.3)	5.9 (5.9-5.9)	320.8 (320.4-321.3)
Ethnicity				
White	83.8 (82.0-85.6)	302.3 (281.1-323.5)	85.7 (85.7-85.7)	289.7 (289.6-289.8)
Black or African American	10.6 (9.2-12.0)	224.3 (192.2-256.4)	10.7 (10.7-10.7)	293.7 (293.3-294.0)
Other race	5.6 (4.4-6.8)	158.1 (123.8-192.4)	3.6 (3.6-3.6)	272.7 (272.2-273.3)
Geographic region				
Northeast	20.8 (18.1-23.5)	332.0 (285.4-378.6)	21.7 (21.7-21.7)	292.0 (291.8-292.3)
Midwest	21.2 (18.5-23.9)	279.6 (242.8-316.4)	16.3 (16.3-16.4)[d]	286.4 (286.1-286.6)
South	36 (32.5-39.5)	264.7 (229.8-299.6)	43.1 (43.1-43.2)[c]	281.7 (281.5-281.8)
West	22 (19.3-24.7)	257.6 (221.7-293.5)	18.8 (18.8-18.8)[d]	283.4 (283.2-283.7)

[a] Pairwise comparisons for the equivalence of percent distributions between National Ambulatory Medical Care Surveys and Centricity Electronic Medical Record were conducted using the two one-sided test. The unmarked categories were all equivalent at a 2.5 percentage point margin.

[b] N/A: not applicable.

[c] Not equivalent at both 2.5 and 5.0 percentage point margins.

[d] Not equivalent at 2.5 percentage point margin.

Looking into office visits' percent distribution by age, it was similar between data sources (*P*=.22 for overall and *P*=.23 by age and sex). The CEMR contains fewer visits by younger patients: the age group <15 years did not reach equivalence at the 5 percentage point equivalence margin for overall comparison and was not equivalent at the 2.5 percentage point margin in comparisons by sex. Age groups of 15-24/25-44 years were equally likely to have a visit with proportions of 7.4% (95% CI 6.6-8.2)/19.3% (95% CI 17.5-21.1) and 7.5% (95% CI 7.4-7.5)/20.0% (95% CI 20.0-20.0) in the NAMCS and the CEMR, respectively. Overall, there were 277.9 (95% CI 259.3-296.5) and 284.6 (95% CI 284.4-284.7) office visits per 100 persons in 2016 in the NAMCS and the CEMR, respectively. Younger patients had similar numbers of visits per year per 100 persons: 257.4 (95% CI 205.5-309.3) and 276.5 (95% CI 276.2-276.8) in the NAMCS and the CEMR, respectively; middle age groups (15-44 years) had significantly fewer visits in the NAMCS; and patients older than 65 years had significantly more visits in the NAMCS compared with the CEMR (*P*<.05).

The overall ethnicity distribution was similar between the NAMCS and CEMR groups (*P*=.20). The proportion of visits by white among all visits were similar in the CEMR (85.7% [95% CI 85.7-85.7]) and NAMCS (83.8% [95% CI 82.0-85.6]; equivalent at 2.5 percentage point margin). The number of visits per 100 persons per year was also similar: CEMR, 289.7 (95% CI 289.6-289.8); and NAMCS, 302.3 (95% CI 281.1-323.5). Although the share of office visits by black or African Americans was similar in both data sources (11%, equivalent at 2.5 percentage point margin), there were significantly fewer office visits per 100 persons per year in NAMCS compared with CEMR in this ethnic group: 224.3 (95% CI 192.2-256.4) versus 293.7 (95% CI 293.3-294.0); *P*<.05.

The geographical distribution of office locations in the CEMR and the NAMCS was similar (Table 1; *P*=.23), with underrepresented Midwest and West (not equivalent at 2.5 percentage point margin) and overrepresented South in the CEMR compared with the NAMCS (not equivalent at 5 percentage point margin).

Prevalence of Chronic Conditions

Obesity

Compared with the NHANES 2015-2016 report, the total obesity prevalence in adults was similar: 39.6% (95% CI 36.5-42.7) in the NHANES and 42.1% (95% CI 42.0-42.1) in the CEMR (Table 2; equivalent at 2.5 percentage point margin). Subgroup analyses revealed a lower proportion of obese males in the NHANES compared with the CEMR: 37.9% (95% CI 33.4-42.4) versus 42.7% (95% CI 42.7-42.8; not equivalent at 2.5 percentage point margin, equivalent at 5 percentage point margin), with the poorest agreement between males aged 40 to 59 years (not equivalent at 7.5 percentage point margin).

Hypertension

During the years 2015-2016, hypertension prevalence in adults was higher in the CEMR than in the NHANES: 33.5% (95% CI 33.5-33.5) versus 29.0% (95% CI 27.0-31.0; Table 2; not equivalent at 2.5 percentage point margin and equivalent at 5 percentage point margin). However, in the CEMR, the prevalence was significantly lower for older patients (aged 60+ years), not equivalent at 7.5 percentage point margin.

High Total Cholesterol

The proportions of adults with high total cholesterol were similar across the NHANES and the CEMR, with a total of 12.4% (Table 2). Although the 95% CI of the proportion of males with high total cholesterol indicated a significant difference between the NHANES (95% CI 9.6%-13.2%) and the CEMR (95% CI 9.3%-9.3%), the proportions appeared to be equivalent at 2.5 percentage point margin based on the TOST.

Diabetes

In the NHANES, the proportion of adults with diabetes during the years 2013-2016 was reported to be 14.0% (95% CI 12.8-15.2); among them, 9.7% (95% CI 8.7%-10.7%) were diagnosed and 4.3% (95% CI 3.5%-5.1%) were undiagnosed (Table 3). In the CEMR, 10.1% of adults were estimated to have diabetes, 8.1% of adults had a diagnostic code, and 2% of adults were without (false negatives). Comparing total diabetes estimates in the CEMR with *diagnosed* in the NHANES, the total and by gender prevalence was similar in both data sources (equivalent at 2.5 percentage point margin), and there were fewer seniors (aged 60+ years) with estimated diabetes in the CEMR compared with the NHANES: 16.4% (95% CI 16.4-16.4) versus 21.0% (95% CI 18.5-23.5; not equivalent at 2.5 percentage point margin and equivalent at 5 percentage point margin).

Table 2. Prevalence of chronic conditions in adult populations in the National Health and Nutrition Examination Surveys and the Centricity Electronic Medical Record.

Characteristics[a]	National Health and Nutrition Examination Surveys 2015-2016			Centricity Electronic Medical Record 2015-2016		
	Total, % (95% CI)	Male, % (95% CI)	Female, % (95% CI)	Total, % (95% CI)	Male, % (95% CI)	Female, % (95% CI)
Obesity						
All adults	39.6 (36.5-42.7)	37.9 (33.4-42.4)	41.1 (38.0-44.2)	42.1 (42.0-42.1)	42.7 (42.7-42.8)[b]	41.5 (41.5-41.6)
Adults aged 20-39 years	35.7 (32.0-39.4)	34.8 (29.3-40.3)	36.5 (33.4-39.6)	34.8 (34.8-34.8)	34.5 (34.5-34.6)	35.0 (35.0-35.0)
Adults aged 40-59 years	42.8 (37.7-47.9)	40.8 (35.1-46.5)	44.7 (38.6-50.8)	47.9 (47.9-47.9)[c]	49.4 (49.4-49.5)[d]	46.7 (46.7-46.7)
Adults aged 60+ years	41.0 (37.3-44.7)	38.5 (35.0-42.0)	43.1 (37.6-48.6)	41.2 (41.2-41.2)	41.6 (41.6-41.6)[b]	40.9 (40.9-40.9)
Hypertension						
All adults	29.0 (27.0-31.0)	30.2 (27.5-32.9)	27.7 (25.7-29.7)	33.5 (33.5-33.5)[b]	36.1 (36.0-36.1)[c]	31.6 (31.6-31.7)[b]
Adults aged 20-39 years	7.5 (5.5-9.5)	9.2 (6.5-11.9)	5.6 (3.4-7.8)	9.6 (9.6-9.6)	11.3 (11.2-11.3)	8.6 (8.6-8.6)[b]
Adults aged 40-59 years	33.2 (29.9-36.5)	37.2 (31.5-42.9)	29.4 (25.5-33.3)	30.6 (30.6-30.7)[b]	33.4 (33.4-33.5)[b]	28.6 (28.5-28.6)
Adults aged 60+ years	63.1 (59.0-67.2)	58.5 (54.2-62.8)	66.8 (61.7-71.9)	52.2 (52.2-52.2)[d]	53.4 (53.4-53.4)[d]	51.3 (51.3-51.3)[d]
High total cholesterol						
All adults	12.4 (10.7-14.1)	11.4 (9.6-13.2)	13.2 (11.0-15.4)	12.4 (12.3-12.4)	9.3 (9.3-9.3)	14.8 (14.8-14.8)
Adults aged 20-39 years	7.9 (6.4-9.4)	9.1 (7.2-11.0)	6.7 (4.2-9.2)	7.4 (7.3-7.4)	9.5 (9.5-9.5)	5.8 (5.8-5.9)
Adults aged 40-59 years	17.1 (14.1-20.1)	16.5 (12.8-20.2)	17.7 (14.9-20.5)	15.1 (15.1-15.1)	12.9 (12.9-12.9)[b]	17.0 (17.0-17.0)
Adults aged 60+ years	12.5 (11.1-13.9)	6.9 (4.3-9.5)	17.2 (14.4-20.0)	11.9 (11.9-11.9)	6.1 (6.1-6.1)	16.6 (16.6-16.6)

[a] Pairwise comparisons for the equivalence of percent distributions between National Ambulatory Medical Care Surveys and the Centricity Electronic Medical Record were conducted using the two one-sided test. The unmarked categories were all equivalent at a 2.5 percentage point margin.

[b] Not equivalent at 2.5 percentage point margins.

[c] Not equivalent at 5.0 percentage point margins.

[d] Not equivalent at 7.5 percentage point margins.

Table 3. The prevalence of diabetes in adult populations.

Characteristics	National Health and Nutrition Examination Surveys 2013-2016			Centricity Electronic Medical Record 2013-2016		
	Total, % (95% CI)	Diagnosed, % (95% CI)	Undiagnosed, % (95% CI)	Total, % (95% CI)	Coded, % (95% CI)	Noncoded, % (95% CI)
All	14.0 (12.8-15.2)	9.7 (8.7-10.7)	4.3 (3.5-5.1)	10.1 (10.1-10.1)	8.1 (8.1-8.1)	2.0 (2.0-2.0)
Men	15.9 (14.1-17.7)	10.8 (9.1-12.5)	5.1 (3.7-6.5)	11.5 (11.5-11.5)	9.2 (9.2-9.2)	1.8 (1.8-1.8)
Women	12.2 (10.7-13.7)	8.8 (7.5-10.1)	3.4 (2.6-4.2)	9.1 (9.1-9.1)	7.3 (7.3-7.3)	2.3 (2.3-2.3)
Adults aged 20-39 years	3.5 (2.2-4.8)	1.8 (1.1-2.5)	1.7 (0.9-2.5)	2.5 (2.5-2.5)	2.1 (2.1-2.1)	0.4 (0.4-0.4)
Adults aged 40-59 years	16.3 (13.9-18.7)	11.1 (9.1-13.1)	5.2 (3.6-6.8)	9.4 (9.4-9.4)	7.7 (7.7-7.7)	1.7 (1.7-1.7)
Adults aged 60+ years	28.2 (25.3-31.1)	21.0 (18.5-23.5)	7.2 (5.7-8.7)	16.4 (16.4-16.5)	12.9 (12.9-12.9)	3.5 (3.5-3.5)

Discussion

Principal Findings

In this study, we compared the CEMR ambulatory and primary care database with federal reports based on the NAMCS and the NHANES. Although the CEMR and the CDC reports may not be directly compared because of the differences in the data collection nature and methodologies applied, in this study, we have observed that the CEMR is a good source of population health research with regard to cardiometabolic conditions. Specifically, we observed that (1) on average, there were 3 office visits per patient per year in both the NAMCS and the CEMR; (2) the distribution of age at office visits in the CEMR is biased toward older population; (3) although the proportional share of all visits by males and females were similar, females/males had more/fewer visits in the NAMCS, compared with the CEMR; (4) the distribution of office visits were similar for ethnic groups; (5) West regions are underrepresented and South region is overrepresented in the CEMR compared with the NAMCS; (6) compared with the CDC reports based on the NHANES data, the prevalence of obesity and high total cholesterol is similar in the CEMR, whereas hypertension prevalence is 5% higher; and (7) the prevalence of diabetes in the CEMR reflects the diagnosed US population well.

A decade ago, Crawford et al [45] compared CEMR's office visits during the year 2005 with the NAMCS report. Crawford et al [45] reported that the CEMR had higher proportions of visits by younger patients and by females, compared with the NAMCS. Maintaining similar methods, we observed a reversed trend in age and no difference in the distribution of visits by gender. The overall prevalence in adults with obesity, high cholesterol, and diabetes was similar between the CEMR and the NHANES reports, and the proportion of those with hypertension was higher in the CEMR than in the NHANES. A possible explanation for this result is the ability to track longitudinal information in EMRs, which is especially prominent in chronic conditions. This feature of EMRs provides exceptional opportunities in terms of extending and modifying the classical epidemiologic theory [57].

Closed systems in the West (the Kaiser Permanente [58]) may explain the finding of lower rates for office visits in the West region in the CEMR. Although the CDC reports were adjusted and weighted with the US Census data, CEMR's crude prevalence estimates of chronic conditions reported in this study are thus biased by geographic regions. This issue should be carefully taken care of in future population health research based on the CEMR data.

Strengths and Limitations

As with any survey, the results in the NAMCS and the NHANES are subject to sampling and nonsampling errors. Nonsampling errors include reporting and processing errors as well as biases because of nonresponse and incomplete response. In the NAMCS, ethnicity data were missing for 25% of visits [47,48], whereas in the CEMR, unknown race accounted for only 9% of visits.

It is important to highlight that the population that seeks medical care is biased from the general population, and healthy individuals will be underrepresented in any EMR database. For this reason, rather than comparing the CEMR with US Census data, we compared the demographics at the time of office visits in the CEMR with the NAMCS, which is carefully weighted and adjusted for the US population. Although a significant subset of CEMR users is participating in the Medical Quality Improvement Consortium, the CEMR research database is biased toward these practices. The participation and response rates in the NAMCS of less than 40% also introduce a selection bias, although the NAMCS estimates were corrected [47,48].

The limitations of this study include the nonavailability of provider specialty and insurance data in the CEMR for deeper comparisons with the NAMCS. Certain specialties might adopt EMRs at a slower rate, and insured individuals might be overrepresented in commercial databases, compared with the general population. Owing to large cohort sizes in the CEMR, reported CIs are very narrow, and we believe they do not reflect meaningful differences. Adopting TOST with 2.5, 5.0, and 7.5 percentage point equivalence margins for data comparison provides another overview of the equivalence of data sources. As mentioned earlier, CEMR and survey methods are not directly comparable, and we have done our best to match the data as closely as possible. The cardiometabolic prevalence estimates should be interpreted carefully, in the light of methodological and regional differences. However, we believe that the results of this study demonstrate the ability of the CEMR to reflect population health quite well.

Conclusions

To conclude, epidemiological and population health findings based on the CEMR database might reflect trends in the general US population; however, the possible region, age, and gender biases presented in this study should be treated and interpreted carefully.

Acknowledgments

Melbourne EpiCentre gratefully acknowledges the support from the National Health and Medical Research Council and the Australian Government's National Collaborative Research Infrastructure Strategy initiative through Therapeutic Innovation Australia.

Authors' Contributions

OM and SP were responsible for the primary design of the study. OM and JD conducted the data extraction and analyses. The manuscript was developed by OM and SP. SP had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. No separate funding was obtained for this study.

Conflicts of Interest

SP has acted as a consultant and/or speaker for Novartis, GI Dynamics, Roche, AstraZeneca, Guangzhou Zhongyi Pharmaceutical, and Amylin Pharmaceuticals LLC. He has received grants from the investigator and investigator-initiated clinical studies from Merck, Novo Nordisk, AstraZeneca, Hospira, Amylin Pharmaceuticals, Sanofi-Avensis, and Pfizer. OM and JD declare no competing interests.

References

1. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, Innovative Medicines Initiative 2nd Programme, Big Data for Better Outcomes, BigData@Heart Consortium of 20 Academic and Industry Partners including ESC. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. Eur Heart J 2018 Apr 21;39(16):1481-1495 [FREE Full text] [doi: 10.1093/eurheartj/ehx487] [Medline: 29370377]
2. Birkhead GS. Successes and continued challenges of electronic health records for chronic disease surveillance. Am J Public Health 2017 Sep;107(9):1365-1367. [doi: 10.2105/AJPH.2017.303938] [Medline: 28787206]
3. Rassen JA, Bartels DB, Schneeweiss S, Patrick AR, Murk W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. Clin Epidemiol 2019;11:1-15 [FREE Full text] [doi: 10.2147/CLEP.S181242] [Medline: 30588119]
4. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015 Mar 18;36:345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
5. Robbins T, Keung SN, Sankar S, Randeva H, Arvanitis TN. Diabetes and the direct secondary use of electronic health records: using routinely collected and stored data to drive research and understanding. Digit Health 2018;4:2055207618804650 [FREE Full text] [doi: 10.1177/2055207618804650] [Medline: 30305917]
6. Hecht J. The future of electronic health records. Nature 2019 Sep;573(7775):S114-S116. [doi: 10.1038/d41586-019-02876-y] [Medline: 31554991]
7. Sepper R, Ross P, Tiik M. Nationwide health data management system: a novel approach for integrating biomarker measurements with comprehensive health records in large populations studies. J Proteome Res 2011 Jan 7;10(1):97-100. [doi: 10.1021/pr1007784] [Medline: 21080730]
8. Blak B, Thompson M, Dattani H, Bourke A. Generalisability of the health improvement network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care 2011;19(4):251-255 [FREE Full text] [doi: 10.14236/jhi.v19i4.820] [Medline: 22828580]
9. Schmidt M, Schmidt SA, Adelborg K, Sundbøll J, Laugesen K, Ehrenstein V, et al. The Danish health care system and epidemiological research: from health care contacts to database records. Clin Epidemiol 2019;11:563-591 [FREE Full text] [doi: 10.2147/CLEP.S179083] [Medline: 31372058]
10. Kosiborod M, Cavender MA, Fu AZ, Wilding JP, Khunti K, Holl RW, CVD-REAL Investigators and Study Group. Lower risk of heart failure and death in patients initiated on sodium-glucose cotransporter-2 inhibitors versus other glucose-lowering drugs: the CVD-real study (comparative effectiveness of cardiovascular outcomes in new users of sodium-glucose cotransporter-2 inhibitors). Circulation 2017 Jul 18;136(3):249-259 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.117.029190] [Medline: 28522450]
11. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. Pharmacoepidemiol Drug Saf 2007 Apr;16(4):393-401. [doi: 10.1002/pds.1335] [Medline: 17066486]
12. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in the health improvement network. Pharmacoepidemiol Drug Saf 2009 Aug;18(8):730-736. [doi: 10.1002/pds.1774] [Medline: 19479713]
13. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
14. Brixner D, Said Q, Kirkness C, Oberg B, Ben-Joseph R, Oderda G. Assessment of cardiometabolic risk factors in a national primary care electronic health record database. Value Health 2007 Jan;10:S29-S36. [doi: 10.1111/j.1524-4733.2006.00152.x]
15. Adamson D, Chang S, Hansen L. Patient Privacy Rights. 2005. Health Research Data for the Real World: The MarketScan Databases URL: http://patientprivacyrights.org/wp-content/uploads/2011/06/Thomson-Medstat-white-paper.pdf [accessed 2020-05-23]

16. Kulaylat A, Schaefer E, Messaris E, Hollenbeak C. Truven health analytics marketscan databases for clinical research in colon and rectal surgery. Clin Colon Rectal Surg 2019 Jan;32(1):54-60 [FREE Full text] [doi: 10.1055/s-0038-1673354] [Medline: 30647546]
17. Gellad WF, Good CB, Lowe JC, Donohue JM. Variation in prescription use and spending for lipid-lowering and diabetes medications in the veterans affairs healthcare system. Am J Manag Care 2010 Oct;16(10):741-750 [FREE Full text] [Medline: 20964470]
18. Bohnert KM, Ilgen MA, Louzon S, McCarthy JF, Katz IR. Substance use disorders and the risk of suicide mortality among men and women in the US veterans health administration. Addiction 2017 Jul;112(7):1193-1201. [doi: 10.1111/add.13774] [Medline: 28301070]
19. Asche CV, Kim J, Kulkarni AS, Chakravarti P, Andersson K. Assessment of association of increased heart rates to cardiovascular events among healthy subjects in the United States: analysis of a primary care electronic medical records database. ISRN Cardiol 2011;2011:924343 [FREE Full text] [doi: 10.5402/2011/924343] [Medline: 22347663]
20. Montvida O, Cai X, Paul SK. Cardiovascular risk factor burden in people with incident type 2 diabetes in the US receiving antidiabetic and cardioprotective therapies. Diabetes Care 2019 Apr;42(4):644-650. [doi: 10.2337/dc18-1865] [Medline: 30679305]
21. Unni S, Wittbrodt E, Ma J, Schauerhamer M, Hurd J, Ruiz-Negrón N, et al. Comparative effectiveness of once-weekly glucagon-like peptide-1 receptor agonists with regard to 6-month glycaemic control and weight outcomes in patients with type 2 diabetes. Diabetes Obes Metab 2018 Feb;20(2):468-473. [doi: 10.1111/dom.13107] [Medline: 28862808]
22. Inzucchi SE, Tunceli K, Qiu Y, Rajpathak S, Brodovicz KG, Engel SS, et al. Progression to insulin therapy among patients with type 2 diabetes treated with sitagliptin or sulphonylurea plus metformin dual therapy. Diabetes Obes Metab 2015 Oct;17(10):956-964 [FREE Full text] [doi: 10.1111/dom.12489] [Medline: 25962401]
23. Levin P, Wei W, Miao R, Ye F, Xie L, Baser O, et al. Therapeutically interchangeable? A study of real-world outcomes associated with switching basal insulin analogues among US patients with type 2 diabetes mellitus using electronic medical records data. Diabetes Obes Metab 2015 Mar;17(3):245-253 [FREE Full text] [doi: 10.1111/dom.12407] [Medline: 25359227]
24. Chitnis AS, Ganz ML, Benjamin N, Langer J, Hammer M. Clinical effectiveness of liraglutide across body mass index in patients with type 2 diabetes in the United States: a retrospective cohort study. Adv Ther 2014 Sep;31(9):986-999 [FREE Full text] [doi: 10.1007/s12325-014-0153-5] [Medline: 25245811]
25. Davis KL, Tangirala M, Meyers JL, Wei W. Real-world comparative outcomes of US type 2 diabetes patients initiating analog basal insulin therapy. Curr Med Res Opin 2013 Sep;29(9):1083-1091. [doi: 10.1185/03007995.2013.811403] [Medline: 23734906]
26. Paul SK, Shaw JE, Montvida O, Klein K. Weight gain in insulin-treated patients by body mass index category at treatment initiation: new evidence from real-world data in patients with type 2 diabetes. Diabetes Obes Metab 2016 Dec;18(12):1244-1252. [doi: 10.1111/dom.12761] [Medline: 27502528]
27. Ma X, Steensma DP, Scott BL, Kiselev P, Sugrue MM, Swern AS. Selection of patients with myelodysplastic syndromes from a large electronic medical records database and a study of the use of disease-modifying therapy in the United States. BMJ Open 2018 Jul 23;8(7):e019955 [FREE Full text] [doi: 10.1136/bmjopen-2017-019955] [Medline: 30037860]
28. Brixner DI, McAdam-Marx C, Ye X, Lau H, Munger MA. Assessment of time to follow-up visits in newly-treated hypertensive patients using an electronic medical record database. Curr Med Res Opin 2010 Aug;26(8):1881-1891. [doi: 10.1185/03007995.2010.489785] [Medline: 20528221]
29. Ashton V, Zhang Q, Zhang NJ, Zhao C, Ramey DR, Neff D, et al. LDL-C levels in US patients at high cardiovascular risk receiving rosuvastatin monotherapy. Clin Ther 2014 May;36(5):792-799. [doi: 10.1016/j.clinthera.2014.03.010] [Medline: 24768187]
30. Chopra I, Kamal KM. Factors associated with therapeutic goal attainment in patients with concomitant hypertension and dyslipidemia. Hosp Pract (1995) 2014 Apr;42(2):77-88. [doi: 10.3810/hp.2014.04.1106] [Medline: 24769787]
31. Saseen JJ, Ghushchyan V, Kaila S, Allen RR, Nair KV. Maintaining goal blood pressures after switching from olmesartan to other angiotensin receptor blockers. J Clin Hypertens (Greenwich) 2013 Dec;15(12):888-892 [FREE Full text] [doi: 10.1111/jch.12197] [Medline: 24102728]
32. Crawford AG, Cote C, Couto J, Daskiran M, Gunnarsson C, Haas K, et al. Prevalence of obesity, type II diabetes mellitus, hyperlipidemia, and hypertension in the United States: findings from the GE centricity electronic medical record database. Popul Health Manag 2010 Jun;13(3):151-161. [doi: 10.1089/pop.2009.0039] [Medline: 20521902]
33. Brixner D, Bron M, Bellows B, Ye X, Yu J, Raparla S, et al. Evaluation of cardiovascular risk factors, events, and costs across four BMI categories. Obesity (Silver Spring) 2013 Jun;21(6):1284-1292 [FREE Full text] [doi: 10.1002/oby.20215] [Medline: 23913737]
34. der Sarkissian M, Bhak RH, Huang J, Buchs S, Vekeman F, Smolarz BG, et al. Maintenance of weight loss or stability in subjects with obesity: a retrospective longitudinal analysis of a real-world population. Curr Med Res Opin 2017 Jun;33(6):1105-1110. [doi: 10.1080/03007995.2017.1307173] [Medline: 28294635]
35. Paul SK, Montvida O, Best JH, Gale S, Pethoe-Schramm A, Sarsour K. Effectiveness of biologic and non-biologic antirheumatic drugs on anaemia markers in 153,788 patients with rheumatoid arthritis: new evidence from real-world data.

Semin Arthritis Rheum 2018 Feb;47(4):478-484 [FREE Full text] [doi: 10.1016/j.semarthrit.2017.08.001] [Medline: 28947313]

36. Rajagopalan V, Alemao E, Kawabata H, Solomon D. SAT0069 performance of the Framingham cardiovascular risk prediction model with and without c-reactive protein or erythrocyte sedimentation rate in RA: analysis of us electronic medical records database. In: Proceedings of the Poster Presentations: Rheumatoid Arthritis - Prognosis, Predictors and Outcome. 2014 Presented at: EULAR'14; June 11-14, 2014; Paris, France. [doi: 10.1136/annrheumdis-2014-eular.1833]
37. Wang J, Mullins CD, Mamdani M, Rublee DA, Shaya FT. New diagnosis of hypertension among celecoxib and nonselective NSAID users: a population-based cohort study. Ann Pharmacother 2007 Jun;41(6):937-943. [doi: 10.1345/aph.1H659] [Medline: 17488830]
38. Tandon N, Carter C, Haas S, Gunnarsson C. Psy64 Ge Centricity Electronic Medical Records Study: Comorbidities and Biologic Experience Among Patients Receiving Golimumab. In: Proceedings of the Poster Session I Disease-Specific Study: Systemic Disorders/Conditions. 2011 Presented at: ISPOR'11; May 21-25, 2016; Washington, DC, USA. [doi: 10.1016/j.jval.2011.02.396]
39. Patel A, Chan W, Aparasu RR, Ochoa-Perez M, Sherer JT, Medhekar R, et al. Effect of psychopharmacotherapy on body mass index among children and adolescents with bipolar disorders. J Child Adolesc Psychopharmacol 2017 May;27(4):349-358. [doi: 10.1089/cap.2016.0133] [Medline: 28422528]
40. Asche C, Said Q, Joish V, Hall CO, Brixner D. Assessment of COPD-related outcomes via a national electronic medical record database. Int J Chron Obstruct Pulmon Dis 2008;3(2):323-326 [FREE Full text] [doi: 10.2147/copd.s1857] [Medline: 18686742]
41. Patel A, Medhekar R, Ochoa-Perez M, Aparasu RR, Chan W, Sherer JT, et al. Care provision and prescribing practices of physicians treating children and adolescents with ADHD. Psychiatr Serv 2017 Jul 1;68(7):681-688. [doi: 10.1176/appi.ps.201600130] [Medline: 28196459]
42. Marelli C, Gunnarsson C, Ross S, Haas S, Stroup DF, Cload P, et al. Statins and risk of cancer: a retrospective cohort analysis of 45,857 matched pairs from an electronic medical records database of 11 million adult Americans. J Am Coll Cardiol 2011 Jul 26;58(5):530-537 [FREE Full text] [doi: 10.1016/j.jacc.2011.04.015] [Medline: 21777752]
43. Talal AH, LaFleur J, Hoop R, Pandya P, Martin P, Jacobson I, et al. Absolute and relative contraindications to pegylated-interferon or ribavirin in the US general patient population with chronic hepatitis C: results from a US database of over 45 000 HCV-infected, evaluated patients. Aliment Pharmacol Ther 2013 Feb;37(4):473-481 [FREE Full text] [doi: 10.1111/apt.12200] [Medline: 23289640]
44. Montvida O, Shaw J, Atherton JJ, Stringer F, Paul SK. Long-term trends in antidiabetes drug usage in the US: real-world evidence in patients newly diagnosed with type 2 diabetes. Diabetes Care 2018 Jan;41(1):69-78. [doi: 10.2337/dc17-1414] [Medline: 29109299]
45. Crawford AG, Cote C, Couto J, Daskiran M, Gunnarsson C, Haas K, et al. Comparison of GE centricity electronic medical record database and national ambulatory medical care survey findings on the prevalence of major conditions in the United States. Popul Health Manag 2010 Jun;13(3):139-150. [doi: 10.1089/pop.2009.0036] [Medline: 20568974]
46. GE Healthcare Systems. 2011. Centricity Electronic Medical Record: Experience that counts URL: http://www3.gehealthcare.com/~/media/Downloads/us/Product/Product-Categories/Healthcare%20IT/Electronic%20Medical%20Records/ITP01981010ENUScentricityemrbrochure.pdf?Parent=%7B3FB3AC2B-38EE-4838-B06E-4472AFC090F0%7D [accessed 2020-05-23]
47. Centers for Disease Control and Prevention. 2016. National Hospital Ambulatory Medical Care Survey: 2016 Emergency Department Summary Tables URL: https://www.cdc.gov/nchs/data/nhamcs/web_tables/2016_ed_web_tables.pdf [accessed 2020-05-23]
48. Centers for Disease Control and Prevention. 2019. About the Ambulatory Health Care Surveys: National Ambulatory Medical Care Survey URL: https://www.cdc.gov/nchs/ahcd/about_ahcd.htm [accessed 2020-05-23]
49. Centers for Disease Control and Prevention. About the National Health and Nutrition Examination Survey: Introduction URL: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm [accessed 2020-05-23]
50. Hales C, Carroll M, Fryar C, Ogden C. Prevalence of obesity among adults and youth: United States, 2015-2016. NCHS Data Brief 2017 Oct(288):1-8 [FREE Full text] [Medline: 29155689]
51. Fryar C, Ostchega Y, Hales CM, Zhang G, Kruszon-Moran D. Hypertension prevalence and control among adults: United States, 2015-2016. NCHS Data Brief 2017 Oct(289):1-8 [FREE Full text] [Medline: 29155682]
52. Carroll M, Fryar CD, Nguyen DT. Total and high-density lipoprotein cholesterol in adults: United States, 2015-2016. NCHS Data Brief 2017 Oct(290):1-8 [FREE Full text] [Medline: 29155686]
53. Mendola N, Chen TC, Gu Q, Eberhardt MS, Saydah S. Prevalence of total, diagnosed, and undiagnosed diabetes among adults: United States, 2013-2016. NCHS Data Brief 2018 Sep(319):1-8 [FREE Full text] [Medline: 30248004]
54. Owusu Adjah ES, Montvida O, Agbeve J, Paul SK. Data mining approach to identify disease cohorts from primary care electronic medical records: a case of diabetes mellitus. Open Bioinforma J 2017 Dec 12;10(1):16-27. [doi: 10.2174/1875036201710010016]

55. McVeigh KH, Newton-Dame R, Chan PY, Thorpe LE, Schreibstein L, Tatem KS, et al. Can electronic health records be used for population health surveillance? Validating population health metrics against established survey data. EGEMS (Wash DC) 2016;4(1):1267 [FREE Full text] [doi: 10.13063/2327-9214.1267] [Medline: 28154837]
56. Tatem KS, Romo ML, McVeigh KH, Chan PY, Lurie-Moroni E, Thorpe LE, et al. Comparing prevalence estimates from population-based surveys to inform surveillance using electronic health records. Prev Chronic Dis 2017 Jun 8;14:E44 [FREE Full text] [doi: 10.5888/pcd14.160516] [Medline: 28595032]
57. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61-81 [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
58. Koebnick C, Langer-Gould AM, Gould MK, Chao CR, Iyer RL, Smith N, et al. Sociodemographic characteristics of members of a large, integrated health care system: comparison with US census bureau data. Perm J 2012;16(3):37-41 [FREE Full text] [doi: 10.7812/tpp/12-031] [Medline: 23012597]

Abbreviations

CDC: Centers for Disease Control and Prevention
CEMR: Centricity Electronic Medical Record
DBP: diastolic blood pressure
EMR: electronic medical record
HbA1c: hemoglobin A1c
NAMCS: National Ambulatory Medical Care Surveys
NHANES: National Health and Nutrition Examination Surveys
PRF: patient record form
RCT: randomized clinical trial
RWD: real-world data
SBP: systolic blood pressure
TOST: two one-sided test

Edited by G Eysenbach; submitted 24.11.19; peer-reviewed by D Gunasekeran, B Dixon, C Fincham; comments to author 17.12.19; revised version received 08.02.20; accepted 21.04.20; published 03.06.20.

Please cite as:
Montvida O, Dibato JE, Paul S
Evaluating the Representativeness of US Centricity Electronic Medical Records With Reports From the Centers for Disease Control and Prevention: Comparative Study on Office Visits and Cardiometabolic Conditions
JMIR Med Inform 2020;8(6):e17174
URL: https://medinform.jmir.org/2020/6/e17174
doi:10.2196/17174
PMID:32490850

Original Paper

Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches

Adane Tarekegn[1], MSc; Fulvio Ricceri[2,3], PhD; Giuseppe Costa[2,3], MPH, MD; Elisa Ferracin[3], BSc; Mario Giacobini[4], PhD

[1]Modeling and Data Science, Department of Mathematics, University of Turin, Turin, Italy

[2]Department of Clinical and Biological Sciences, University of Turin, Turin, Italy

[3]Unit of Epidemiology, Regional Health Service, Local Health Unit Torino 3, Turin, Italy

[4]Data Analysis and Modeling Unit, Department of Veterinary Sciences, University of Turin, Turin, Italy

Corresponding Author:
Adane Tarekegn, MSc
Modeling and Data Science, Department of Mathematics
University of Turin
Via Carlo Alberto, 10
Turin,
Italy
Phone: 39 3394246167
Email: adanenega.tarekegn@unito.it

Abstract

Background: Frailty is one of the most critical age-related conditions in older adults. It is often recognized as a syndrome of physiological decline in late life, characterized by a marked vulnerability to adverse health outcomes. A clear operational definition of frailty, however, has not been agreed so far. There is a wide range of studies on the detection of frailty and their association with mortality. Several of these studies have focused on the possible risk factors associated with frailty in the elderly population while predicting who will be at increased risk of frailty is still overlooked in clinical settings.

Objective: The objective of our study was to develop predictive models for frailty conditions in older people using different machine learning methods based on a database of clinical characteristics and socioeconomic factors.

Methods: An administrative health database containing 1,095,612 elderly people aged 65 or older with 58 input variables and 6 output variables was used. We first identify and define six problems/outputs as surrogates of frailty. We then resolve the imbalanced nature of the data through resampling process and a comparative study between the different machine learning (ML) algorithms – Artificial neural network (ANN), Genetic programming (GP), Support vector machines (SVM), Random Forest (RF), Logistic regression (LR) and Decision tree (DT) – was carried out. The performance of each model was evaluated using a separate unseen dataset.

Results: Predicting mortality outcome has shown higher performance with ANN (TPR 0.81, TNR 0.76, accuracy 0.78, F1-score 0.79) and SVM (TPR 0.77, TNR 0.80, accuracy 0.79, F1-score 0.78) than predicting the other outcomes. On average, over the six problems, the DT classifier has shown the lowest accuracy, while other models (GP, LR, RF, ANN, and SVM) performed better. All models have shown lower accuracy in predicting an event of an emergency admission with red code than predicting fracture and disability. In predicting urgent hospitalization, only SVM achieved better performance (TPR 0.75, TNR 0.77, accuracy 0.73, F1-score 0.76) with the 10-fold cross validation compared with other models in all evaluation metrics.

Conclusions: We developed machine learning models for predicting frailty conditions (mortality, urgent hospitalization, disability, fracture, and emergency admission). The results show that the prediction performance of machine learning models significantly varies from problem to problem in terms of different evaluation metrics. Through further improvement, the model that performs better can be used as a base for developing decision-support tools to improve early identification and prediction of frail older adults.

(JMIR Med Inform 2020;8(6):e16678) doi:10.2196/16678

KEYWORDS

predictive modeling; frailty; machine learning; genetic programming; imbalanced dataset; elderly people; classification

Introduction

Health challenges associated with aging are a major medical and social concern as the burden of the older population is increasing dramatically. The elderly population, which has been conventionally defined as having a chronological age of 65 years or older [1], is becoming a meaningful challenge for every nation in terms of services and costs [2]. According to a 2017 United Nations report [3], the world population of older persons aged 60 years and above was 600 million in 2000, and it is projected to rise to approximately 2 billion by 2050. The aging of the population has profound consequences, with one of the main issues associated with this phenomenon being the higher prevalence of frailty condition [4]. Frailty is one of the most important and emerging age-related conditions that generally represents an increasing limitation in daily activities. Older people develop a wide variety of age-related conditions that contribute to an increase in their vulnerability to minor stressor events and lead to loss of autonomy. This phenomenon is commonly known as frailty [2,5]. People who are considered frail are particularly vulnerable to undesirable outcomes, including disability, injurious falls, hospitalization, and death. These health outcomes result in a poor quality of life and increased demand for medical and social care and are associated with increased costs for individuals and health systems. According to a study [6], health spending increases significantly in higher age classes compared with lower age groups. Older adults (those aged 70 years and older) are more likely to live with multiple chronic conditions and functional limitations. This combination is related to a larger probability of accessing an emergency department (ED) along with higher Medicare spending for inpatient hospitals, trained nursing facilities, and home health services. However, frailty is not an inevitable consequence of aging, and it can be prevented and managed to foster a longer and healthier life. Early detection and screening would help to deliver preventive interventions and reverse frailty conditions.

Several scales and models have been proposed for the detection of frailty [7-10]; however, a precise operational definition of frailty or a standard method for its screening and diagnosis is still lacking [11,12]. In clinical settings where the standard measure of frailty is missing and the care of the elderly is a priority, it is imperative to have a specific model in the prediction of frailty according to the characteristics of the population being studied. Therefore, this study aimed to detect multiple outcomes of frailty (mortality, disability, fracture, hospitalizations, and emergency admissions) using large administrative health databases on elderly people in Piedmont, Italy.

The study examines the existing machine learning techniques (artificial neural networks [ANNs], genetic programming [GP], support vector machines [SVMs], logistic regression [LR], decision trees [DTs] and random forests [RFs]) to predict frailty according to the different adverse health outcomes. These approaches were considered for their performance and practical usefulness in the analysis of different types of medical data.

Methods

Data Source

This study was based on the Piedmontese Longitudinal Study. The data were collected using an individual record linkage available for about 4 million Piedmont (Italy) inhabitants between the Italian 2011 census and the administrative and health databases (enrollees registry, hospital discharges, drug prescriptions, outpatient clinical investigation database, and health exemptions) that is included in the Italian Statistical National Plan. Subjects aged 65 years and above are included in the study. The dataset contains 1,095,612 subjects and 64 variables (58 input and 6 output variables). The dataset includes a wide variety of predictor variables, including clinical and socioeconomic aspects, and six target variables for every subject: mortality, disability, urgent hospitalization, fracture, preventable hospitalization, and accessing the emergency department (ED) with red code. Color codes assigned to patients may vary from one hospital to another, but in this study, a red code is used to identify patients with severe symptoms who need an immediate care. Since we intend to develop predictive models for these frailty indicators, we extracted as input data those collected in 2016, while using as output values those collected in 2017.

For simple implementation and analysis, the data were transformed into six datasets, one for each output variable. As a result, six problems associated with frailty conditions were identified and defined. The six datasets were considered separately in the analysis, which resulted in six independent binary classification problems. All the input variables used in the study are presented in Multimedia Appendix 1. Table 1 contains descriptive statistics for all output variables with the frequency distributions of each category of an output variable represented as counts and percentages. Table 1 clearly shows how the dataset is, for each output variable, unbalanced. In fact, approximately 4% of the records have mortality risk as 1, and the other 96% have mortality risk as 0. There are similar numbers of records having risk as 1 for emergency admission with red code, fracture, preventable hospitalization, disability, and urgent hospitalization. This is clearly an indication of an imbalanced dataset, as the number of subjects from the positive sample is much smaller than the number of subjects of the negative sample.

Table 1. Description of output variables in the dataset.

Variable	Code	Value, n (%)
Mortality		
No	0	1,053,790 (96.18)
Yes	1	41,823 (3.82)
Accessing the ED[a] with red code		
No	0	1,088,124 (99.32)
Yes	1	7489 (0.68)
Disability		
No	0	1,064,186 (97.13)
Yes	1	31,427 (2.87)
Fracture		
No	0	1,088,530 (99.35)
Yes	1	7083 (0.65)
Urgent hospitalization		
No	0	1,056,695 (96.45)
Yes	1	38,918 (3.55)
Preventable hospitalization		
No	0	1,076,541 (98.26)
Yes	1	19,072 (1.74)

[a] ED: emergency department.

Most machine learning techniques suffer from such extremely unbalanced datasets, and, as a result, they may be biased toward the majority class. Instructing a model with an algorithm that tries to maximize the accuracy will naturally lead to classifying everything as the major class and will not give acceptable results.

Handling Imbalanced Dataset

The dataset in each problem (mortality, accessing ED with red code, disability, fracture, urgent hospitalization, and preventable hospitalization) is imbalanced, as shown in Table 1. The imbalanced proportions between the positive and negative classes of the six datasets are treated independently. There are various approaches to deal with imbalanced data that have been used in the literature, such as resampling [13] and cost-sensitive learning methods [14].

In this study, we chose the resampling methods, which are based on undersampling [15] and oversampling [16]. These methods are advantageous because they are classifier independent and can be used as a preprocessing step, in which the processed data can be given as input to any classifier. Oversampling is the process of replicating samples from the minority class to balance the data. The limitation of oversampling is that it may cause an overfitting problem as it clones the same instance and requires more time to execute compared with the undersampling approach. As a result, it is recommended when the dataset is quite small in size. Another issue with oversampling is that as our aim was to detect minority classes, oversampling changes the class that we want to identify, which may not be acceptable in some critical real-time problems [17]. Undersampling balances the imbalanced data by reducing the size of samples from the majority class. One limitation of the undersampling approach is that it may lead to loss of important information or introduce bias in the data. From a practical point of view, some literature showed that undersampling tends to outperform oversampling in some settings [18], while others demonstrate that oversampling performs better than undersampling [19]. In high-dimensional data, oversampling performs worse [20], while undersampling performs worse in very small datasets. In our case, since the amount of collected data is sufficient, we adopted undersampling to rebalance the sample distribution followed by a statistical test to avoid bias and ensure representativeness between samples. Since we have multioutput data, we followed these simple steps to obtain balanced and independent datasets:

- Filter all positive and negative samples from the original dataset based on the values of the output variables. Samples with at least one positive class value from the six outcomes are grouped as a positive sample, which accounts for 10% of the original dataset, and all remaining are grouped as a negative sample, comprising 90% of the original dataset.
- Keeping all the 10% samples in the positive class (minority group), we randomly selected an equal number of samples (10%) from the negative class (majority group).
- Check whether the randomly selected 10% negative samples were representative of the remaining negative samples (90%). After checking that the test was reasonably significant, we obtained a new multioutput dataset of size 211,924 each. A statistical test was applied in all variables to decide whether the distribution of frequencies of a

variable in the 10% sample was representative of the same variable in the 90% sample. Since all variables in the study are categorical, we used a chi-square independence test with a significance level of .05 to check if there was a significant difference between the 10% sample and the 90% sample with respect to input variables. The yielded chi-square statistic and *P* values were assessed to support the significance of the test's conclusion. The results of the chi-square test between 10% and 90% negative samples are shown in Multimedia Appendix 1.

- Once the test was significant, we decomposed the multioutput dataset into six independent datasets. An equal number of positive and negative samples were then selected randomly from each dataset.

Predictive Models

The machine learning approaches selected for this study are SVMs, ANNs, RFs, DTs, LR, and GP. We have presented below a brief summary of these learning algorithms.

SVM is a robust classifier to identify two classes that require a huge amount of training data to select an effective decision boundary. Several studies have used SVM on disease prediction [21-24]. The SVM algorithm is used to predict events by plotting the training dataset where a hyperplane classifies the points into two classes, presence and absence of frailty. SVM is based on kernel functions, which project linearly inseparable input data to higher dimensional space for better classification. Various kernels and parameters are used to improve the performance of classification by SVM [25]. In this study, the radial basis function kernel is used with different values of gamma and the regularization parameters for solving each classification problem.

ANNs are analytical techniques that have been successful in solving classification problems in different domains [26-30]. Based on the functioning of biological neural networks, ANNs are dense networks of interconnected artificial neurons that get activated based on inputs. The multilayer perceptron neural network (MLPNN), one of the most used paradigms in ANNs, is employed in this study. The MLPNN includes one input layer, one or more hidden layers, and one output layer. In MLPNN, the input nodes pass values to the first hidden layer, and the nodes of the first hidden layer pass values to the second layer and so on, until producing outputs. The main parameters used in MLPNN, including activation function, solver, hidden layer size, and learning rate, are configured for each classification work.

We also explored the potential of tree-based classifiers (DTs and RFs) for the prediction of outcomes in each frailty problem. DTs build classification models in the form of a tree structure [31]. The main algorithms used in DTs are ID3, C4.5, and the classification and regression tree [32], which build DTs using the concept of information entropy. In our study, the classification and regression tree algorithm is used for building the DT with hyperparameters set for each problem. RFs consist of a large number of individual DTs that operate as an ensemble. Each tree gives a classification, and the forest chooses the classification having the most votes (over all the trees in the forest). RF is known for the prediction task in the medical domain [33-35]. The hyperparameters (such as the number of trees in the forest, maximum number of features considered for splitting a node, maximum number of levels in each DT, etc) have been set for each problem.

LR, a specific type of multivariate regression, is the most common and well-established binary classifier [36]. LR is used to model only a dichotomous variable, which usually represents the presence or absence of an outcome or event based on a set of predictor variables. It predicts an event of occurrence by fitting a dataset into a logit function. In this study, like other machine learning models, LR has been used to distinguish frail and nonfrail subjects.

Another technique applied to the prediction task is GP, typically designed to address the problem of automatic program synthesis and automatic programming. GP accomplishes this task by generating a population of computer programs over many generations using operations of natural selection [37]. Many works in GP focus on classifier induction, a task that can be accomplished by evolution using GP [38,39]. In GP, setting the control parameters is an important first step to manipulate data and obtain good results. In our datasets, we tried several experiments for classification tasks by using the control parameters of GP proposed in HeuristicLab tools [40]. The parameter values of GP used for our experiment are listed in Multimedia Appendix 2.

Performance Metrics

The performance measures were considered based on the proportion of older people with mortality, urgent hospitalization, preventable hospitalization, disability, fracture, and ED admission with a red code. Predicting these adverse outcomes among a large number of subjects is important when applied in real-world practice. Hence, the true positive rate (TPR) was the main metric to consider. The overall accuracy, true negative rate (TNR), and F1-score, which is the harmonic mean of precision and recall, were used as additional performance metrics. The accuracy, TPR, and TNR were formulated using the true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs). These measures are defined in the equations in Figure 1 [41].

Figure 1. Evaluation metrics.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}}$$

$$\text{True Positive Rate } = \frac{TP}{TP + FN}$$

$$\text{True Negative Rate } = \frac{TN}{TN + FP}$$

$$\text{F1} - \text{Score } = 2 * \frac{precision * recall}{precision + recall}$$

Data Analysis Tools

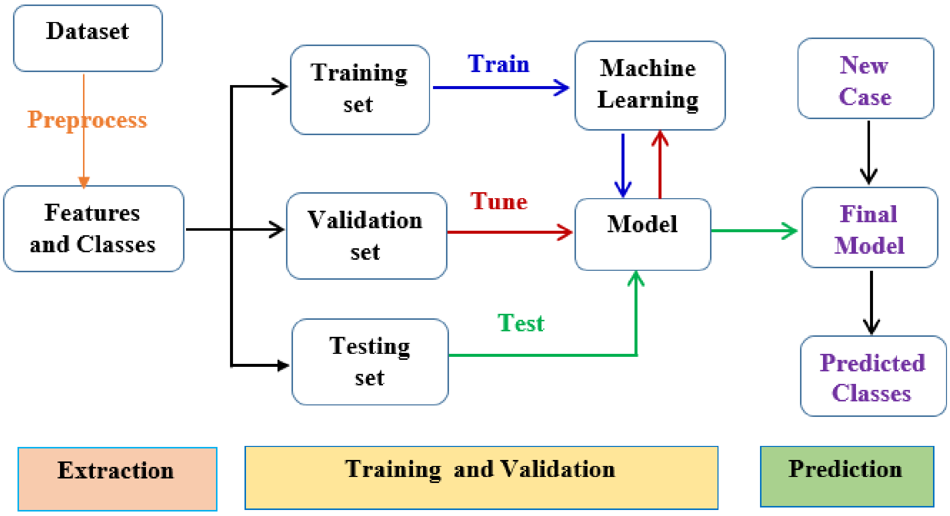
The data analysis tools used in the study are Python Scikit-learn library, RStudio software package, and HeuristicLab. In this work, the exploratory data analysis part and statistical test analysis were done using R3.5.0, whereas the entire classification problems with SVMs, RFs, NNs, and DTs were implemented using Python 3.7. Multimedia Appendix 3 presents some Python codes used in the experiment. HeuristicLab is a software tool for heuristic and evolutionary algorithms. In this study, HeuristicLab was used to carry out classification problems using GP.

Experimental Settings

Model Evaluation

In analyzing the data for prediction, the output variables represent an occurrence in the next year, and the predictive model is proposed to predict frailty according to the expected risk of urgent hospitalization, preventive hospitalization, disability, fracture, accessing the ED with a red code, and death within a year. The performance of various predictive models is evaluated for each outcome prediction using four metrics: accuracy, TPR, TNR, and F1-score. These metrics provide an effective and simple way to evaluate the performance of a classifier. Using these four measures, the models were evaluated using both the holdout method [42] and the cross-validation method [43]. Figure 2 shows the general experimental workflow of the predictive machine learning model.

Figure 2. Experimental workflow of the predictive machine learning model.



Holdout Method

In this study, our first experiment was started by exploring the predictive performance of machine learning methods using the holdout method. This method randomly splits a dataset into training and testing according to a given proportion. Each machine learning model was trained using the training dataset (70%) and evaluated using test datasets (30%). The training dataset was used for building the model, while the test dataset was used to evaluate the prediction capabilities of models.

K-Fold Cross-Validation

The K-fold cross-validation procedure was applied to each problem's data. Cross-validation is one of the most commonly used model evaluation procedure that extends the holdout method by repeating the splitting process several times. The K-fold cross-validation technique divides the dataset into K folds of roughly equal size. The model being evaluated is then trained using the K-1 parts, and one part is left out for model validation. In this study, we used 10-folds, and the dataset was split into three parts for the purpose of model training and testing: the training set to build the model, validation set to select the model parameters, and test set to evaluate the performance of the final model based on the selected parameters.

Hyperparameter Tuning

In all experiments, the set of hyperparameters was selected for each machine learning method before the training begins. Hyperparameters allow machine learning algorithms to better adjust to the problem details. The hyperparameters for each model were tuned using a grid search with cross-validation in Python Scikit-learn as described by Mueller and Guido [44]. Multimedia Appendix 2 presents the list of hyperparameters used for training each machine learning model in this study.

Results

Study Population

From the original dataset of 1,095,612 elderly people aged 65 years and above, we retrieved 83,646 with mortality, 77,836 with urgent hospitalization, 62,854 with disability, 38,144 with preventable hospitalization, 14,978 who accessed the ED with red code, and 14,166 with a fracture for this study. The retrieval process was made using the resampling approach, and each problem was analyzed independently of the others using the widely used machine learning models. In this section, the predictive performance of machine learning models using both holdout and cross-validation methods are presented through feature selection analysis.

Feature Selection

Feature selection provides an effective way to remove irrelevant and/or redundant features, which can reduce running time, increase learning accuracy, and facilitate a better understanding of the model [45,46]. Unnecessary features can also increase the chance of overfitting and decrease the generalization performance on the test data. We used a filter method for feature selection [47,48]. A chi-square test is a filter method used in this study to determine the statistical significance between features and the target. The chi-square value, together with *P* values at a significance level of .05, was used to identify the most important features with their rank (ie, variables shown to be significantly associated with the outcome by the chi-square test analysis [*P*<.05] were selected for model building). *P*<.001 indicates that there is an association between the input and the target variables. The strength of the association between the input variables and the target is ranked based on the chi-square value. Out of the 58 predictor variables, 25, 24, 10, 7, 4, and 3 nonsignificant variables were discarded for preventable hospitalization, urgent hospitalization, emergency admission with red code, fracture, mortality, and disability, respectively. Table 2 presents the top 15 ranked features in order of decreasing importance in the mortality and fracture problems. The most significant feature for other problems is presented in Multimedia Appendix 4.

Table 2. The most important variables in the mortality and fracture problems.

Rank	Mortality problem		Fracture problem	
	Variable	*P* value	Variable	*P* value
1	Age	<.001	Age	<.001
2	Charlson index	<.001	Femur fracture	<.001
3	# urgent hospitalization	<.001	# urgent hospitalization	<.001
4	# total hospitalization	<.001	Neck fracture	<.001
5	Invalidity	<.001	Green code	<.001
6	# nontraumatic	<.001	# total hospitalization	<.001
7	Disability	<.001	Charlson index	<.001
8	Poly prescriptions	<.001	Poly prescriptions	<.001
9	Green code	<.001	Invalidity	<.001
10	Yellow code	<.001	Disability	<.001
11	Blood	<.001	Nerve disease	<.001
12	Anemia	<.001	Depression	<.001
13	Circulatory disease	<.001	Blood	<.001
14	Respiratory disease	<.001	Anemia	<.001
15	Urinary tract disease	<.001	Yellow code	<.001

Feature importance can give us insight into a problem by indicating what variables are the most discriminating between classes. For example, in Table 2, age and the Charlson index are the most important features in the prediction of mortality, which makes sense in the problem context. The rank of features differs from one problem to another, except for the variable age, which has the highest score in all problems. Next to the age attribute, variables such as femur fracture, number of urgent hospitalizations, and neck fracture are the most discriminant features in the fracture problem, while type of family and home living status are the least significant variables. Mental disease, poly prescription, and disease of the circulatory system are variables with the highest rank in urgent hospitalization and preventable hospitalization. The age, Charlson index, and number of urgent hospitalizations are the most important predictors of emergency admission with red code. Some features with the lowest rank and common to urgent hospitalization and preventable hospitalization include marital status, level of education, work status, and income. Each of the predictive models (SVM, ANN, LR, RF, and DT) have been applied using the most important features in each of the six problems. GP differs from the other machine learning models in that it performs implicit feature selection automatically during the evolutionary process. GP learns which combination of features are useful for classification and determines the optimal number of features automatically.

Performance via Holdout Method

In this study, our first experimental results were obtained through the holdout (train-test split) method with all subsets of features (from top 3 to top 58 features) using the default parameters of the models. However, these approaches have brought the problem of overfitting on the training data for RF and DT, as shown in Figures 3 and 4. In order to reduce the overfitting problem and improve performance, the parameters of each model were tuned using grid search along with the most important features associated with each outcome. Table 3 shows the performance of SVM, RF, ANN, DT, and GP using the best features and parameters selected on each problem.

Figure 3. Train accuracy (left) and test accuracy (right) for mortality data without performing any parameter tuning and using all the feature subsets (from top 3 to top 58 feature subsets). The left plot shows that random forest and decision tree overfit the training data, which poorly generalize on the test data as the number of features increase.

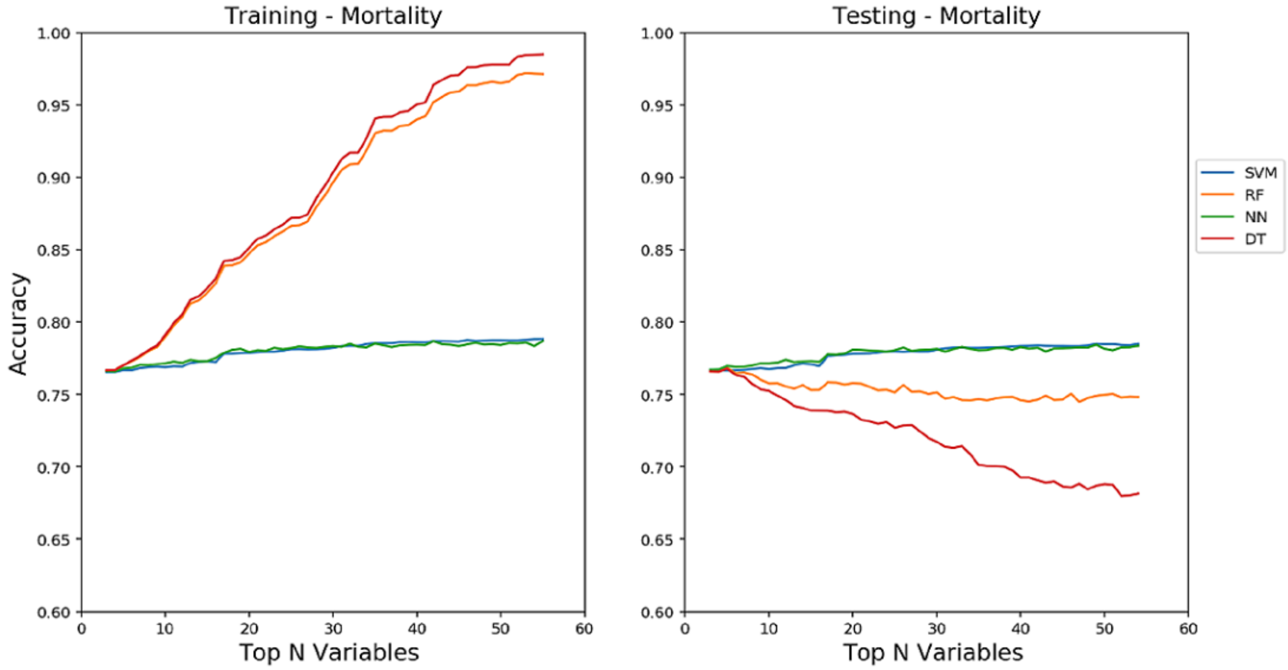


Figure 4. Train accuracy (left) and test accuracy (right) for fracture data without performing parameter tuning and using all the feature subsets (from top 3 to top 58 feature subsets). The left plot shows that random forest and decision tree overfit the training data, which poorly generalize on the test data as the number of features increase.

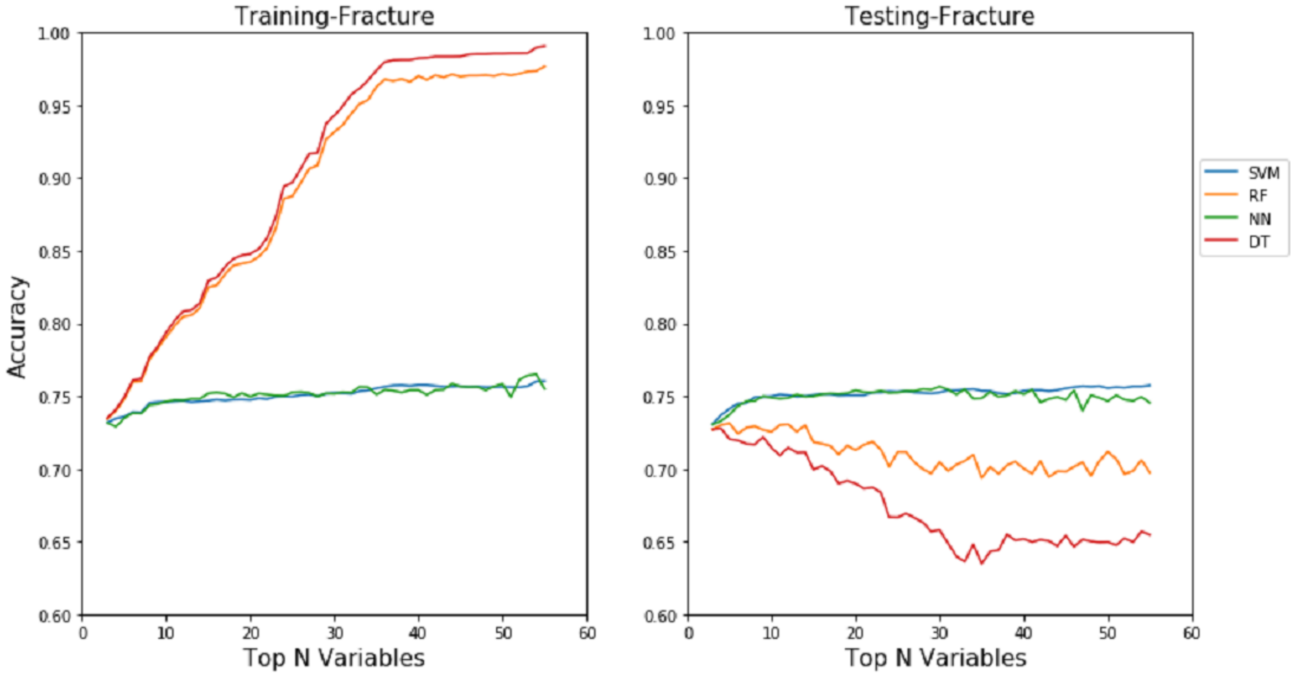


Table 3. Prediction performance using true positive rate and true negative rate for the six problems.

Problem	SVM[a]		RF[b]		ANN[c]		DT[d]		GP[e]	
	TPR[f]	TNR[g]	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR
Mortality	0.78	0.78	0.79	0.77	0.79	0.78	0.60	0.79	0.75	0.76
Disability	0.78	0.72	0.78	0.71	0.75	0.75	0.78	0.69	0.71	0.67
Fracture	0.75	0.74	0.77	0.72	0.77	0.72	0.79	0.66	0.70	0.73
Urgent hospitalization	0.61	0.73	0.65	0.68	0.66	0.68	0.64	0.68	0.66	0.62
Preventable hospitalization	0.74	0.73	0.73	0.72	0.73	0.73	0.76	0.66	0.73	0.64
ED admission[h,i]	0.63	0.73	0.63	0.72	0.63	0.74	0.62	0.73	0.73	0.63

[a] SVM: support vector machine.

[b] RF: random forest.

[c] ANN: artificial neural network.

[d] DT: decision tree.

[e] GP: genetic programming.

[f] TPR: true positive rate.

[g] TNR: true negative rate.

[h] ED: emergency department.

[i] with a red code.

In our experiments, we explored common variations for each machine learning algorithm in frailty predictions. From the results of the experiment in Table 3, it is clear that all algorithms behave differently for each different problem. For the mortality dataset, RF and ANN produced higher values of TPR (0.79) while the DT produced the lowest performance. For the fracture problem, DT scored the highest values of TPR (0.79), while GP scored the lowest value. GP, on the other hand, has higher values of TPR on the urgent hospitalization dataset. The overall average TPR of RF was slightly higher for all problems, while SVM has slightly higher values of TNR in all problems and DT produced the lowest average TPR in all problems. According to the results on the test part of the dataset, all machine learning models showed lower prediction performance on the urgent hospitalization and accessing the ED with red code problems, while mortality and disability have higher values of prediction results compared with other outcomes. On the disability problem, GP has lower TPR compared with SVM, RF, ANN, and DT, while it has the highest TPR on accessing the ED with red code. For other problems, GP produces comparable results. The performance of GP is compared with other machine learning methods using statistical tests to draw better conclusions. We performed a pairwise statistical test between the 30 runs of GP and each individual machine learning model using the Wilcoxon signed-rank test. The Wilcoxon statistical test is a nonparametric test that ranks the differences in performances of GP and other algorithms over each frailty problem. The Wilcoxon test is based on the TPR of each algorithm in each problem on the test data. The results of the test in terms of *P* values with the significance level of .01 are shown in Table 4.

Table 4. Results of Wilcoxon signed-rank test in terms of *P* values.

Problem/dataset	SVM[a] vs GP[b]	RF[c] vs GP	NN[d] vs GP	DT[e] vs GP
Mortality	<.001	.003	.001	<.001
Fracture	<.001	.02	<.001	.002
Disability	.06	.004	.01	.003
Urgent hospitalization	.71	.01	.37	.01
Preventable hospitalization	.68	.03	.87	.005
Accessing the ED[f] with a red code	.006	<.001	.01	<.001

[a] SVM: support vector machine.

[b] GP: genetic programming.

[c] RF: random forest.

[d] NN: neural network.

[e] DT: decision tree.

[f] ED: emergency department.

As depicted in Table 4, the Wilcoxon test allows rejecting 11 hypotheses. The *P* values below .01 indicate that the respective algorithms differ significantly in TPR, while the *P* values above .01 indicate that the algorithms behave similarly in predicting frailty conditions. The test results between SM and GP are statistically significant only in disability, urgent hospitalization, and preventable hospitalization. Combining the experimental results and Wilcoxon signed-rank test results, it is concluded that for mortality and fracture SVM outperformed GP in the TPR score, while GP outperformed SVM and RF on urgent hospitalization and accessing the ED with red code. Despite the fact that DT represented higher values of TPR on the preventable hospitalization compared with other algorithms, its lowest TNR result represented a higher disadvantage. ANN has a similar performance with GP for preventable and urgent hospitalization events.

Performance via 10-Fold Cross-Validation

The 10-fold cross-validation reduces the variance of the resulting estimate by averaging over 10 different subsamples. This 10-fold cross-validation can deal with limitations of the holdout method, such as to reduce overfitting, and therefore is more reliable and provides better generalization performance on the test data. Thus, in our second experiment, we used the 10-fold cross-validation method on each of the six datasets. The variation of each model's accuracy across the 10 samples in the 10-fold cross-validation is presented in Figures 5 and 6 for the largest dataset (ie, mortality) and smallest dataset (ie, fracture), respectively. From the figures, one can see that the models are more stable in predicting mortality than fracture across the 10 samples. It is also found a slight variation of classification rate across the 10 samples for the other outcomes.

Figure 5. The score of five models across 10 validation samples on the mortality problem.

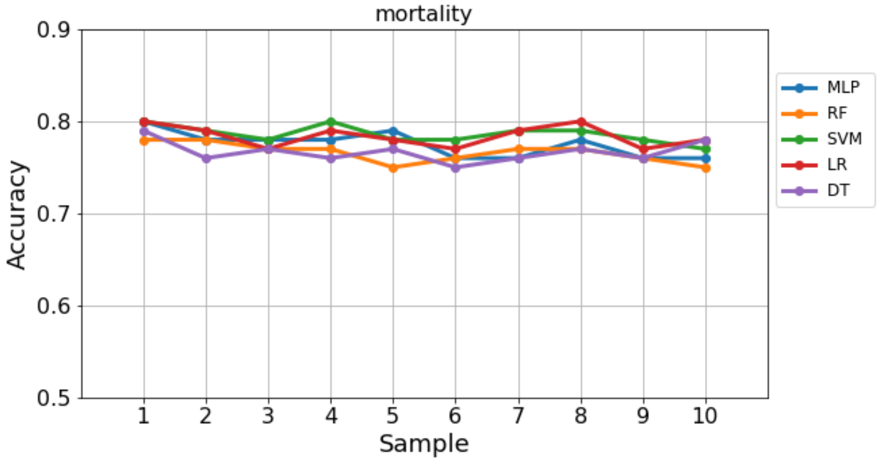
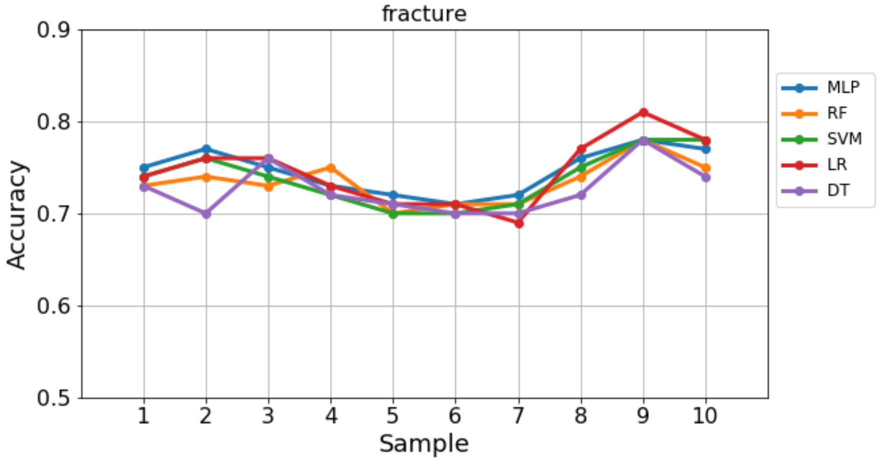


Figure 6. The score of five models across 10 validation samples on the fracture problem.



As shown from Figure 5, the classification rate across 10 samples in the 10-fold cross-validation is slightly varied in each classifier for the mortality problem. The variation of accuracy is greater in the fracture problem from sample 1 to sample 10 for each model, as shown in Figure 6. Particularly, LR has shown the greatest variation of performance among the models, where it performed the lowest accuracy at sample 7 and the highest accuracy at sample 9 in the fracture problem. DT has shown the highest classification rate at sample 10 for mortality and at sample 3 in the fracture problem, while it has the lowest accuracy in the rest of the samples. The average performance of 10-fold cross-validation in each problem is shown in Table 5, where performance for each model is measured using accuracy, TPR, TNR, and F1-score.

Table 5. Prediction results of models using a 10-fold cross-validation.

Models	Accuracy	TPR[a]	TNR[b]	F1-score
Mortality				
ANN[c]	0.78	0.81	0.76	0.79
SVM[d]	0.79	0.77	0.80	0.78
RF[e]	0.78	0.79	0.76	0.76
LR[f]	0.78	0.78	0.79	0.78
DT[g]	0.75	0.80	0.70	0.76
Fracture				
ANN	0.75	0.77	0.73	0.75
SVM	0.75	0.77	0.74	0.75
RF	0.75	0.78	0.72	0.76
LR	0.75	0.75	0.75	0.75
DT	0.74	0.76	0.72	0.74
Disability				
ANN	0.74	0.76	0.71	0.75
SVM	0.75	0.78	0.73	0.76
RF	0.75	0.77	0.72	0.75
LR	0.75	0.76	0.73	0.74
DT	0.73	0.78	0.70	0.75

[a] TPR: true positive rate.

[b] TNR: true negative rate.

[c] ANN: artificial neural network

[d] SVM: support vector machine.

[e] RF: random forest.

[f] LR: logistic regression.

[g] DT: decision tree.

From the results of all models in each outcome presented in Tables 5 and 6, we can see that predicting mortality events has shown the highest performance, while predicting urgent hospitalization and accessing the ED with red code have shown lower performance. Next to the mortality problem, better prediction performance is obtained on disability and fracture problems. This implies that the dataset in this study is better at predicting mortality than predicting the other outcomes. In predicting urgent hospitalization, only SVM achieved the best performing algorithm in all measurements (accuracy, TPR, TNR, and F1-score) among all models trained using 10-fold cross-validation. In the mortality problem, the highest average performance was obtained by ANN (accuracy 0.78, TPR 0.81, TNR 0.76, F1-score 0.79) and SVM (accuracy 0.79, TPR 0.77, TNR 0.80, F1-score 0.78) followed by LR (accuracy 0.78, TPR 0.78, TNR 0.79, F1-score 0.78). DT produced the highest TPR (0.80), and RF showed comparable results (accuracy 0.78, TPR 0.79, TNR 0.76, F1-score 0.76) on the mortality problem. For the fracture and disability problems, SVM, RF, and LR have similar accuracy (0.75), although they all differ in TPR, TNR, and F1-score.

Table 6. Prediction results of models using a 10-fold cross-validation procedure.

Models	Accuracy	TPR[a]	TNR[b]	F1-score
Urgent hospitalization				
ANN[c]	0.67	0.64	0.71	0.66
SVM[d]	0.75	0.77	0.73	0.76
RF[e]	0.66	0.65	0.67	0.66
LR[f]	0.67	0.72	0.62	0.65
DT[g]	0.66	0.65	0.67	0.65
Preventable hospitalization				
ANN	0.74	0.73	0.74	0.73
SVM	0.74	0.71	0.76	0.73
RF	0.73	0.73	0.74	0.73
LR	0.74	0.71	0.76	0.73
DT	0.72	0.73	0.71	0.72
Accessing the ED[h] with red code				
ANN	0.70	0.65	0.74	0.67
SVM	0.68	0.64	0.72	0.66
RF	0.68	0.66	0.70	0.67
LR	0.69	0.64	0.74	0.67
DT	0.67	0.70	0.65	0.68

[a] TPR: true positive rate.

[b] TNR: true negative rate.

[c] ANN: artificial neural network.

[d] SVM: support vector machine.

[e] RF: random forest.

[f] LR: logistic regression.

[g] DT: decision tree.

[h] ED: emergency department.

From the results of the experiments, it is important to observe that the various machine learning techniques can significantly vary in terms of their performance for the different evaluation metrics. For example, in the mortality problem, SVM outperformed DT and ANN in TNR value (.80), and ANN outperformed both SVM and DT in F1-score (0.79), while DT outperformed both models in TPR value (0.80). The performance of all models differs in all problems due to the difference in feature space, size, and diversity of data in each of the six problems. The prediction performance of all models trained with mortality data (largest in size) is much better than the performance of models trained with accessing the ED with red code data (smaller in size), which demonstrates that the size of data is an important factor for better performance, but this is not always true for all models. In addition, the performance of each machine learning technique varied from problem to problem. For example, the performance of ANN measured in TPR is 0.81, 0.77, 0.76, 0.74, 0.70, and 0.67 for mortality, fracture, disability, preventable hospitalization, accessing the ED with red code, and urgent hospitalization, respectively, while for DT the TPR is 0.80, 0.75, 0.78, 0.73, 0.70, and 0.65 for each problem, respectively. Considering the performance of these two machine learning methods (ANN and DT) in their TPR value, ANN outperforms DT in mortality and fracture problems, while DT outperforms ANN in disability and accessing the ED with red code problems. We can also see that LR has a higher TPR value than SVM in the mortality problem. This shows that it is not necessarily true that the more complex machine learning models (eg, ANN, SVM) always outperform simpler models (eg, DT, LR). In 10-fold cross-validation, the RF classifiers achieved comparable performance to SVM and ANN in most of the problems. On the other hand, tree-based classifiers (RF and DT) are more sensitive to bad features and quality of data. Therefore, effective feature selection is an important step to improve their performance. The SVM model tends to perform well in high-dimensional classification problems; however, it may not perform well if the sample classes of the problem are highly overlapping. ANN can generally outperform other techniques if the dataset is very large and if the structure of the dataset is complex (eg, if it has many layers).

In general, machine learning is an exploratory process, where there is no one-size-fits-all problem. In particular, there is no model recognized to achieve supreme performance for all problem types, domains, or datasets [49]. The best performing machine learning model differs from one problem to another according to the characteristics of variables, size of the data, and metrics used. The idea is similar to the "no free lunch" theorem [50,51], which states that there is no universal algorithm that works best for every problem. However, it is important to study each problem by evaluating each model carefully in order to reach an effective predictive design. The results also show it is essential to carefully explore and evaluate the performance of machine learning techniques using various optimized parameter values as well as using the most significant predictor variables. Particularly, tree-based classifiers (eg, RF and DT) are more sensitive to overfitting problems, as shown in Figures 3 and 4 on the mortality and fracture problems, if the correct subset of features is not selected or if the required parameter values of models are not configured properly. The accuracy in the figures clearly indicates that an increasing number of features in RF and DT leads to the model overfitting. Interestingly, SVM and ANN models showed relatively consistent performance both on training and testing even with an increasing number of features.

Discussion

Principal Findings

A predictive model that can use administrative health data will be useful in various settings to classify those individuals who are at risk of frailty and deliver preventive interventions. In this study, we performed several experiments using different classification techniques to build predictive models for frailty. The results show that machine learning models can vary significantly from problem to problem in terms of different evaluation metrics. The explored models have shown solid predictive power to better estimate the risk of mortality than predicting disability, fracture, emergency admission in red code, urgent hospitalization, and preventable hospitalization within the next year. Although each model is not a comprehensive model to predict all frailty outcomes, we have demonstrated that the SVM model has shown higher overall accuracy (0.79) in predicting mortality and urgent hospitalization than other models, when using 10-fold cross-validation. On the other hand, except for the ANN, all other machine learning models have shown relatively poor overall accuracy in predicting emergency admission with red code.

In addition, our results show significant performance enhancement by reducing features. In order to reduce the overfitting problem and improve the prediction performance of classifiers, the feature selection process is executed, where the best subset of the available features is chosen. In each binary classification problem, all independent variables were ranked using the chi-square feature selection method for each outcome in both holdout and cross-validation methods. Using 10-fold cross-validation on mortality problems, the TPR values (also called sensitivity) of ANN, SVM, RF, LR, and DT were 0.81, 0.77, 0.79, 0.78, and 0.80, respectively. In the holdout method, almost similar results were obtained for ANN, SVM, and RF, while DT produced higher TPR values using 10-fold cross-validation than holdout method on the mortality problem. In general, 10-fold cross-validation reduces variance by averaging over 10 different partitions; it is then less sensitive to any of the partitioning bias in the training and testing data. On predicting emergency admission with red code, GP achieved better TPR value than SVM, ANN, LR, RF, and DT, while SVM outperformed all models in predicting urgent hospitalization in all evaluation measures.

Generally, an important observation from the results of the experiments is that on average some of the machine learning models produce quite similar results from the same outcome, while the best performing model varies from one outcome to another outcome in terms of different metrics. For example, SVM and ANN produce similar performance on average across all evaluation metrics in mortality and hospitalization outcomes. RF and LR produced similar performance on average across all measurements in disability and fracture outcomes. However, the prediction results of each machine learning model varies from mortality to fracture or fracture to hospitalization, etc. This can demonstrate the feasibility of identifying frail older subjects through routinely collected administrative health databases.

Strengths and Limitations

The strength of our study is the possibility to include a multidimensional administrative database using the most powerful predictive machine learning models. In contrast to the previous studies, the prediction models use a wide variety of input variables, including clinical and socioeconomic aspects, with six simultaneous outcomes. The use of routinely collected socioclinical data can represent the multidimensional loss of an individual's reserves, which allows predicting prospective outcomes in the elderly. Moreover, the predictions of frailty in terms of the six adverse outcomes were assessed and analyzed, which is a step forward in studying the association of frailty with multiple health conditions on a frail person.

There are limitations to our study. Even though the original data comes with multiple outcomes, each machine learning algorithm was designed to predict a single outcome, and each result is analyzed independently of the others. Therefore, further studies should investigated constructing a predictive model that considers the correlations among the output variables to provide a list of relevant outputs for a given, previously unseen patient. Furthermore, patient information such as gender can be included in the study in order to understand gender-related factors for frailty and their impact on hospitalization and mortality among older people.

Conclusions

Predictive modeling using the information available from administrative health databases is an efficient method to identify frail older people appropriate for interventions to prevent adverse outcomes. The proposed predictive models can be applied to detect and predict frail people who are at increased risk of adverse outcomes. This study suggests that a machine learning–based predictive model could be used to screen future frailty conditions using clinical and socioeconomic variables, which are commonly collected in community health care institutions. With efforts to enhance predictive performance, such a machine learning–based approach can further contribute to the improvement of frailty interventions in the elderly community.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of variables and statistical test between samples.
[DOCX File , 50 KB - medinform_v8i6e16678_app1.docx]

Multimedia Appendix 2
Hyperparameters used for training models.
[DOCX File , 18 KB - medinform_v8i6e16678_app2.docx]

Multimedia Appendix 3
Python implementation codes used in the study.
[RAR File , 305 KB - medinform_v8i6e16678_app3.rar]

Multimedia Appendix 4
List of most important features in each outcome.
[DOCX File , 18 KB - medinform_v8i6e16678_app4.docx]

References

1. Health statistics and information systems: proposed working definition of an older person in Africa for the MDS Project. Geneva: World Health Organization URL: http://www.who.int/healthinfo/survey/ageingdefnolder/en/index.html [accessed 2020-05-18]
2. Kojima G, Liljas AEM, Iliffe S. Frailty syndrome: implications and challenges for health care policy. Risk Manag Healthc Policy 2019;12:23-30 [FREE Full text] [doi: 10.2147/RMHP.S168750] [Medline: 30858741]
3. United Nations. World Population Ageing 2017 Highlights URL: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2017_Highlights.pdf [accessed 2020-05-18]
4. Comans TA, Peel NM, Hubbard RE, Mulligan AD, Gray LC, Scuffham PA. The increase in healthcare costs associated with frailty in older people discharged to a post-acute transition care program. Age Ageing 2016 Mar;45(2):317-320. [doi: 10.1093/ageing/afv196] [Medline: 26769469]
5. Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005 Aug 30;173(5):489-495 [FREE Full text] [doi: 10.1503/cmaj.050051] [Medline: 16129869]
6. Neuman T, Cubanski J, Huang J, Damico A. Kaiser Family Foundation. 2015 Jan 14. The rising cost of living longer: analysis of Medicare spending by age for beneficiaries in traditional Medicare URL: http://files.kff.org/attachment/report-the-rising-cost-of-living-longer-analysis-of-medicare-spending-by-age-for-beneficiaries-in-traditional-medicare [accessed 2020-05-18]
7. Aguayo GA, Donneau A, Vaillant MT, Schritz A, Franco OH, Stranges S, et al. Agreement between 35 published frailty scores in the general population. Am J Epidemiol 2017 Aug 15;186(4):420-434 [FREE Full text] [doi: 10.1093/aje/kwx061] [Medline: 28633404]
8. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci 2001 Mar;56(3):M146-M156. [Medline: 11253156]
9. Maxwell CA, Wang J. Understanding frailty: a nurse's guide. Nurs Clin North Am 2017 Sep;52(3):349-361. [doi: 10.1016/j.cnur.2017.04.003] [Medline: 28779818]
10. Mohd Hamidin FA, Adznam SN, Ibrahim Z, Chan YM, Abdul Aziz NH. Prevalence of frailty syndrome and its associated factors among community-dwelling elderly in East Coast of Peninsular Malaysia. SAGE Open Med 2018;6:2050312118775581 [FREE Full text] [doi: 10.1177/2050312118775581] [Medline: 29872529]
11. Fougère B, Kelaiditi E, Hoogendijk EO, Demougeot L, Duboué M, Vellas B, et al. Frailty index and quality of life in nursing home residents: results from INCUR study. J Gerontol A Biol Sci Med Sci 2016 Mar;71(3):420-424. [doi: 10.1093/gerona/glv098] [Medline: 26297653]
12. Santos-Eggimann B, Sirven N. Screening for frailty: older populations and older individuals. Public Health Rev 2016;37:7 [FREE Full text] [doi: 10.1186/s40985-016-0021-8] [Medline: 29450049]

13. Lee PH. Resampling methods improve the predictive power of modeling in class-imbalanced datasets. Int J Environ Res Public Health 2014 Sep 18;11(9):9776-9789 [FREE Full text] [doi: 10.3390/ijerph110909776] [Medline: 25238271]
14. Khan SH, Hayat M, Bennamoun M, Sohel FA, Togneri R. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans Neural Netw Learn Syst 2018 Aug;29(8):3573-3587. [doi: 10.1109/TNNLS.2017.2732482] [Medline: 28829320]
15. Liu X, Wu J, Zhou Z. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B Cybern 2009 Apr;39(2):539-550. [doi: 10.1109/TSMCB.2008.2007853] [Medline: 19095540]
16. Parsa AB, Taghipour H, Derrible S, Mohammadian AK. Real-time accident detection: coping with imbalanced data. Accid Anal Prev 2019 Aug;129:202-210. [doi: 10.1016/j.aap.2019.05.014] [Medline: 31170559]
17. Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, et al. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access 2016;4:7940-7957. [doi: 10.1109/access.2016.2619719]
18. Wallace B, Small K, Brodley C, Trikalinos T. Class imbalance, redux. 2011 Presented at: IEEE 11th International Conference on Data Mining; 2011; Vancouver. [doi: 10.1109/icdm.2011.33]
19. Naseriparsa M, Mansour Riahi Kashani M. Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset. Int J Comput Appl 2013 Sep 18;77(3):33-38. [doi: 10.5120/13376-0987]
20. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013 Mar 22;14:106 [FREE Full text] [doi: 10.1186/1471-2105-14-106] [Medline: 23522326]
21. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104-116 [FREE Full text] [doi: 10.1016/j.csbj.2016.12.005] [Medline: 28138367]
22. Pan L, Liu G, Mao X, Li H, Zhang J, Liang H, et al. Development of prediction models using machine learning algorithms for girls with suspected central precocious puberty: retrospective study. JMIR Med Inform 2019 Feb 12;7(1):e11728 [FREE Full text] [doi: 10.2196/11728] [Medline: 30747712]
23. Huang M, Chen C, Lin W, Ke S, Tsai C. SVM and SVM ensembles in breast cancer prediction. PLoS One 2017;12(1):e0161501 [FREE Full text] [doi: 10.1371/journal.pone.0161501] [Medline: 28060807]
24. O'Dwyer L, Lamberton F, Bokde ALW, Ewers M, Faluyi YO, Tanner C, et al. Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. PLoS One 2012;7(2):e32441 [FREE Full text] [doi: 10.1371/journal.pone.0032441] [Medline: 22384251]
25. Maji S, Berg AC, Malik J. Efficient classification for additive kernel SVMs. IEEE Trans Pattern Anal Mach Intell 2013 Jan;35(1):66-77. [doi: 10.1109/TPAMI.2012.62] [Medline: 22392703]
26. Woldaregay AZ, Årsand E, Walderhaug S, Albers D, Mamykina L, Botsis T, et al. Data-driven modeling and prediction of blood glucose dynamics: machine learning applications in type 1 diabetes. Artif Intell Med 2019 Jul;98:109-134. [doi: 10.1016/j.artmed.2019.07.007] [Medline: 31383477]
27. Putra FR, Nursetyo AA, Thakur SS, Roy RB, Syed-Abdul S, Malwade S, et al. Prediction of clinical events in hemodialysis patients using an artificial neural network. Stud Health Technol Inform 2019 Aug 21;264:1570-1571. [doi: 10.3233/SHTI190539] [Medline: 31438236]
28. Dihge L, Ohlsson M, Edén P, Bendahl P, Rydén L. Artificial neural network models to predict nodal status in clinically node-negative breast cancer. BMC Cancer 2019 Jun 21;19(1):610 [FREE Full text] [doi: 10.1186/s12885-019-5827-6] [Medline: 31226956]
29. Bradley R, Tagkopoulos I, Kim M, Kokkinos Y, Panagiotakos T, Kennedy J, et al. Predicting early risk of chronic kidney disease in cats using routine clinical laboratory tests and machine learning. J Vet Intern Med 2019 Sep 26 [FREE Full text] [doi: 10.1111/jvim.15623] [Medline: 31557361]
30. Wellner B, Grand J, Canzone E, Coarr M, Brady PW, Simmons J, et al. Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. JMIR Med Inform 2017 Nov 22;5(4):e45 [FREE Full text] [doi: 10.2196/medinform.8680] [Medline: 29167089]
31. Aris-Brosou S, Kim J, Li L, Liu H. Predicting the reasons of customer complaints: a first step toward anticipating quality issues of in vitro diagnostics assays with machine learning. JMIR Med Inform 2018 May 15;6(2):e34 [FREE Full text] [doi: 10.2196/medinform.9960] [Medline: 29764796]
32. Lee J. Patient-specific predictive modeling using random forests: an observational study for the critically ill. JMIR Med Inform 2017 Jan 17;5(1):e3 [FREE Full text] [doi: 10.2196/medinform.6690] [Medline: 28096065]
33. Wu J, Zan X, Gao L, Zhao J, Fan J, Shi H, et al. A machine learning method for identifying lung cancer based on routine blood indices: qualitative feasibility study. JMIR Med Inform 2019 Aug 15;7(3):e13476 [FREE Full text] [doi: 10.2196/13476] [Medline: 31418423]
34. Beunza J, Puertas E, García-Ovejero E, Villalba G, Condes E, Koleva G, et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). J Biomed Inform 2019 Sep;97:103257. [doi: 10.1016/j.jbi.2019.103257] [Medline: 31374261]

35. Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. Proc IEEE Inst Electr Electron Eng 2018 Apr;106(4):690-707 [FREE Full text] [doi: 10.1109/JPROC.2017.2789319] [Medline: 30886441]
36. Hyun S, Moffatt-Bruce S, Cooper C, Hixon B, Kaewprag P. Prediction model for hospital-acquired pressure ulcer development: retrospective cohort study. JMIR Med Inform 2019 Jul 18;7(3):e13785 [FREE Full text] [doi: 10.2196/13785] [Medline: 31322127]
37. Poli R, Langdon W, McPhee N. A Field Guide to Genetic Programming. London: Lulu Enterprises; 2008.
38. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. BioData Min 2011 May 11;4:12 [FREE Full text] [doi: 10.1186/1756-0381-4-12] [Medline: 21569330]
39. Gao W, Chen X, Chen D. Genetic programming approach for predicting service life of tunnel structures subject to chloride-induced corrosion. J Adv Res 2019 Nov;20:141-152 [FREE Full text] [doi: 10.1016/j.jare.2019.07.001] [Medline: 31452958]
40. HeuristicLab. URL: https://dev.heuristiclab.com/trac.fcgi/ [accessed 2019-02-21]
41. Olsen DL. Advanced Data Mining Techniques. New York: Springer; 2008.
42. Wshah S, Skalka C, Price M. Predicting posttraumatic stress disorder risk: a machine learning approach. JMIR Ment Health 2019 Jul 22;6(7):e13946 [FREE Full text] [doi: 10.2196/13946] [Medline: 31333201]
43. Sena GR, Lima TPF, Mello MJG, Thuler LCS, Lima JTO. Developing machine learning algorithms for the prediction of early death in elderly cancer patients: usability study. JMIR Cancer 2019 Sep 26;5(2):e12163 [FREE Full text] [doi: 10.2196/12163] [Medline: 31573896]
44. Müller A, Guido S. Introduction to Machine Learning with Python. Sebastopol: O'Reilly Media; 2015.
45. Suomi V, Komar G, Sainio T, Joronen K, Perheentupa A, Blanco Sequeiros R. Comprehensive feature selection for classifying the treatment outcome of high-intensity ultrasound therapy in uterine fibroids. Sci Rep 2019 Jul 29;9(1):10907 [FREE Full text] [doi: 10.1038/s41598-019-47484-y] [Medline: 31358836]
46. Cerruela García G, Pérez-Parras Toledano J, de Haro García A, García-Pedrajas N. Filter feature selectors in the development of binary QSAR models. SAR QSAR Environ Res 2019 May;30(5):313-345. [doi: 10.1080/1062936X.2019.1588160] [Medline: 31112077]
47. Alirezanejad M, Enayatifar R, Motameni H, Nematzadeh H. Heuristic filter feature selection methods for medical datasets. Genomics 2019 Jul 02. [doi: 10.1016/j.ygeno.2019.07.002] [Medline: 31276753]
48. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: introduction and review. J Biomed Inform 2018 Sep;85:189-203 [FREE Full text] [doi: 10.1016/j.jbi.2018.07.014] [Medline: 30031057]
49. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 2013 Apr;66(4):398-407 [FREE Full text] [doi: 10.1016/j.jclinepi.2012.11.008] [Medline: 23384592]
50. Wolpert D. The lack of a priori distinctions between learning algorithms. Neural Computation 1996;8(7):1341-1390 [FREE Full text] [doi: 10.1162/neco.1996.8.7.1341]
51. Spasic I, Krzeminski D, Corcoran P, Balinsky A. Cohort selection for clinical trials from longitudinal patient records: text mining approach. JMIR Med Inform 2019 Oct 31;7(4):e15980 [FREE Full text] [doi: 10.2196/15980] [Medline: 31674914]

Abbreviations

ANN: artificial neural network
DT: decision tree
ED: emergency department
FN: false negative
FP: false positive
GP: genetic programming
LR: logistic regression
MLPNN: multilayer perceptron neural network
RF: random forest
SVM: support vector machine
TN: true negative
TNR: true negative rate
TP: true positive
TPR: true positive rate

Edited by Z Huang; submitted 15.10.19; peer-reviewed by M Boukhechba, Z Zhang, P Giabbanelli; comments to author 07.12.19; revised version received 07.01.20; accepted 16.02.20; published 04.06.20.

Please cite as:
Tarekegn A, Ricceri F, Costa G, Ferracin E, Giacobini M
Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches
JMIR Med Inform 2020;8(6):e16678
URL: http://medinform.jmir.org/2020/6/e16678/
doi:10.2196/16678
PMID:32442149

Original Paper

Improving Clinical Translation of Machine Learning Approaches Through Clinician-Tailored Visual Displays of Black Box Algorithms: Development and Validation

Shannon Wongvibulsin[1], PhD; Katherine C Wu[2], MD; Scott L Zeger[3], PhD

[1]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, United States

[2]Department of Medicine, Division of Cardiology, Johns Hopkins University School of Medicine, Baltimore, MD, United States

[3]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

Corresponding Author:
Scott L Zeger, PhD
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street
Room E3650
Baltimore, MD, 21205
United States
Phone: 1 410 502 9054
Email: sz@jhu.edu

Abstract

Background: Despite the promise of machine learning (ML) to inform individualized medical care, the clinical utility of ML in medicine has been limited by the minimal interpretability and *black box* nature of these algorithms.

Objective: The study aimed to demonstrate a general and simple framework for generating clinically relevant and interpretable visualizations of *black box* predictions to aid in the clinical translation of ML.

Methods: To obtain improved transparency of ML, simplified models and visual displays can be generated using common methods from clinical practice such as decision trees and effect plots. We illustrated the approach based on postprocessing of ML predictions, in this case random forest predictions, and applied the method to data from the Left Ventricular (LV) Structural Predictors of Sudden Cardiac Death (SCD) Registry for individualized risk prediction of SCD, a leading cause of death.

Results: With the LV Structural Predictors of SCD Registry data, SCD risk predictions are obtained from a random forest algorithm that identifies the most important predictors, nonlinearities, and interactions among a large number of variables while naturally accounting for missing data. The *black box* predictions are postprocessed using classification and regression trees into a clinically relevant and interpretable visualization. The method also quantifies the relative importance of an individual or a combination of predictors. Several risk factors (heart failure hospitalization, cardiac magnetic resonance imaging indices, and serum concentration of systemic inflammation) can be clearly visualized as branch points of a decision tree to discriminate between low-, intermediate-, and high-risk patients.

Conclusions: Through a clinically important example, we illustrate a general and simple approach to increase the clinical translation of ML through clinician-tailored visual displays of results from black box algorithms. We illustrate this general model-agnostic framework by applying it to SCD risk prediction. Although we illustrate the methods using SCD prediction with random forest, the methods presented are applicable more broadly to improving the clinical translation of ML, regardless of the specific ML algorithm or clinical application. As any trained predictive model can be summarized in this manner to a prespecified level of precision, we encourage the use of simplified visual displays as an adjunct to the complex predictive model. Overall, this framework can allow clinicians to peek inside the black box and develop a deeper understanding of the most important features from a model to gain trust in the predictions and confidence in applying them to clinical care.

(JMIR Med Inform 2020;8(6):e15791) doi:10.2196/15791

KEYWORDS

machine learning; interpretability; clinical translation; prediction models; visualization

Introduction

Background

There is growing interest in benefiting from the predictive power of machine learning (ML) to improve the outcomes of medical care at more affordable costs. Although notable for their impressive predictive ability, ML *black box* predictions are often characterized by minimal interpretability, limiting their clinical adoption despite their promise for improving health care [1-6]. As a result, there is growing emphasis on the field of *interpretable ML* or *explainable Artificial Intelligence* to provide explanations of how models make their decisions [6-8]. However, the lack of understanding of how ML predictions are generated and the complex relationships between the predictors and outcomes are still obstacles to the adoption of ML in clinical practice.

Many approaches have been developed to explain predictions and determine ML feature importance and effect, but they have limited adoption in real-world clinical applications [9-12]. There have been previous proposals to stack ML methods or to use rule extraction with ML output to produce simpler summaries, but because of their inherent complexities or lack of clinical applications, these tools are seldom used in medicine [13-16].

Objectives

To accelerate the integration of ML into clinical care, an emphasis on the personalization of these tools for the end user is crucial. Our work is motivated by the well-known clinical challenge of measuring an individual's risk of sudden cardiac death (SCD), a leading cause of death with inherently complex pathophysiology that lends itself to novel approaches [17-22]. Although we focus here on SCD as an illustrative example, the methods we present are applicable more broadly to improving clinical translation of ML, regardless of the specific ML algorithm or clinical application. The contribution of this work is a general framework for translating complex *black box* predictions into easily understood representations through commonly encountered clinical summaries. Overall, we emphasize the need for multidisciplinary teams to create clinician-tailored visual displays that provide interpretability in ways that are personalized to the clinician’s preferences for understanding ML predictions to aid in effective clinical translation of ML.

Methods

Data Source

The Left Ventricular (LV) Structural Predictors of SCD Registry is a prospective observational registry (clinicaltrials.gov, NCT01076660), which enrolled 382 patients for the primary end point of an adjudicated appropriate implantable cardioverter defibrillator firing for ventricular tachycardia or ventricular fibrillation or SCD not aborted by the device [23-29]. In the 8-year follow-up, 75 individuals had the primary outcome.

Modeling

Our ML approach is based on the random forest (RF) algorithm implemented in the randomForestSRC R package [30] extended to time-varying SCD risk prediction [31]. RF is an ensemble learning method based on a collection of decision trees, where the overall RF prediction is the ensemble average or majority vote. Random sampling of predictor variables at each decision tree node and bootstrapping the original training data decrease the correlation among the trees in the forest to allow for impressive predictive performance [32,33]. For our RF, the predictors included demographics, comorbidities, medications, electrophysiologic parameters, laboratory values, LV ejection fraction by echocardiography, and cardiac magnetic resonance (CMR) imaging indices, summarized in Table 1.

Table 1. Patient characteristics in the Left Ventricular Structural Predictors of Sudden Cardiac Death Registry (N=382).

Variables	No. of SCD[a] event (n=307)	Patient with SCD event (n=75)	*P* value[b]
Demographics and clinical characteristics			
Age (years), mean (SD)	57 (13)	57 (12)	.75
Male, n (%)	211 (68.7)	63 (84)	*.01*
Race, n (%)			**.66**
White	200 (65.1)	51 (68)	
African American	99 (32)	21 (28)	
Other	8 (3)	3 (4)	
Body surface area (m^2), mean (SD)	1.98 (0.28)	2.05 (0.28)	.07
Ischemic cardiomyopathy etiology, n (%)	149 (48.5)	44 (59)	.15
Years from incident MI[c] or cardiomyopathy diagnosis, mean (SD)	3.83 (5.18)	5.43 (5.61)	*.02*
NYHA[d] functional class, n (%)			**.55**
I	64 (21)	20 (27)	
II	137 (44.6)	31 (41)	
III	106 (34.5)	24 (32)	
One or more heart failure hospitalizations, n (%)	0 (0)	19 (25.3)	*<.001*
Cardiac risk factors, n (%)			
Hypertension	180 (58.6)	44 (59)	>.99
Hypercholesterolemia	180 (58.6)	45 (60)	.93
Diabetes	85 (28)	19 (25)	.79
Nicotine use	133 (43.3)	44 (59)	*.02*
Medication usage, n (%)			
ACE[e]-inhibitor or ARB[f]	275 (89.6)	66 (88)	.85
Beta-blocker	288 (93.8)	68 (91)	.48
Lipid-lowering	199 (64.8)	56 (75)	.14
Antiarrhythmics (amiodarone)	18 (6)	8 (11)	.22
Diuretics	173 (56.4)	54 (72)	*.02*
Digoxin	50 (16)	16 (21)	.39
Aldosterone inhibitor	80 (26)	21 (28)	.85
Aspirin	215 (70.0)	55 (73)	.67
Electrophysiologic variables			
Prior atrial fibrillation, n (%)	51 (17)	14 (19)	.80
Ventricular rate (bpm), mean (SD)	73 (14)	70 (14)	.06
QRS duration (ms), mean (SD)	118 (31)	122 (27)	.30
Presence of LBBB[g], n (%)	79 (26)	14 (19)	.26
Biventricular ICD[h], n (%)	90 (29)	17 (23)	.31
Laboratory values or biomarkers			
Sodium (mEq/L), mean (SD)	139 (3)	139 (3)	.73
Potassium (mEq/L), mean (SD)	4.26 (0.42)	4.27 (0.39)	.87
Creatinine (mEq/L), mean (SD)	1.07 (0.59)	1.09 (0.33)	.81
eGFR[i] (mL/min/1.73 m^2), mean (SD)	81 (24)	80 (21)	.80

Variables	No. of SCD[a] event (n=307)	Patient with SCD event (n=75)	*P* value[b]
Blood urea nitrogen (mg/dL), mean (SD)	19.62 (8.72)	20.28 (8.33)	.55
Glucose (mg/dL), mean (SD)	120 (53)	113 (34)	.23
Hematocrit (%), mean (SD)	40 (4)	41 (5)	*.03*
hsCRP[j] (µg/mL), mean (SD)	6.89 (12.87)	9.10 (16.29)	.22
NT-proBNP[k] (ng/L), mean (SD)	2704 (6736)	2519 (1902)	.82
IL-6[l] (pg/mL), mean (SD)	3.05 (5.36)	4.32 (6.28)	.12
IL-10[m] (pg/mL), mean (SD)	10.74 (49.67)	13.67 (59.94)	.70
TNF-αRII[n] (pg/mL), mean (SD)	3425 (1700)	3456 (1671)	.90
cTnT[o] (ng/mL), mean (SD)	0.03 (0.08)	0.02 (0.05)	.62
cTnI[p] (ng/mL), mean (SD)	0.10 (0.28)	0.10 (0.25)	.98
CK-MB[q] (ng/mL), mean (SD)	3.94 (5.77)	3.87 (3.86)	.93
Myoglobin (ng/mL), mean (SD)	31.37 (30.80)	37.13 (41.53)	.31
LVEF[r]: NonCMR[s] LVEF (%), mean (SD)	24.2 (7.6)	23.0 (7.4)	.19
CMR structural and functional indices			
LVEF (%), mean (SD)	27.8 (10.3)	25.1 (8.8)	*.04*
LV[t] end-diastolic volume index (ml/m$^{2)}$, mean (SD)	122.3 (39.9)	136.2 (48.4)	*.01*
LV end-systolic volume index (ml/m^2), mean (SD)	91.5 (39.1)	104.3 (45.2)	*.02*
LV mass index (ml/m^2), mean (SD)	75.1 (24.4)	80.3 (21.2)	*.09*
CMR hyperenhancement			
LGE[u] present (%), mean (SD)	176 (66)	56 (86)	*.002*
Gray zone (g), mean (SD)	8.8 (11.6)	13.8 (12.2)	*.002*
Core (g), mean (SD)	12.4 (14.9)	17.7 (15.1)	*.01*
Total scar (g), mean (SD)	21.1 (25.4)	31.3 (25.6)	*.004*

[a] SCD: sudden cardiac death.

[b] *P* values <.05 are italicized.

[c] MI: myocardial infarction.

[d] NYHA: New York Heart Association.

[e] ACE: angiotensin-converting enzyme.

[f] ARB: angiotensin II receptor blocker.

[g] LBBB: left bundle branch block.

[h] ICD: implantable cardioverter defibrillator.

[i] eGFR: estimated glomerular filtration rate.

[j] hsCRP: high-sensitivity C-reactive protein.

[k] NT-proBNP: N-terminal pro-b-type natriuretic peptide.

[l] IL-6: interleukin-6.

[m] IL-10: interleukin-10.

[n] TNF-αRII: tumor necrosis factor alpha R II.

[o] cTnT: cardiac troponin T.

[p] cTnI: cardiac troponin I.

[q] CK-MB: creatine kinase MB.

[r] LVEF: left ventricular ejection fraction.

[s] CMR: cardiac magnetic resonance.

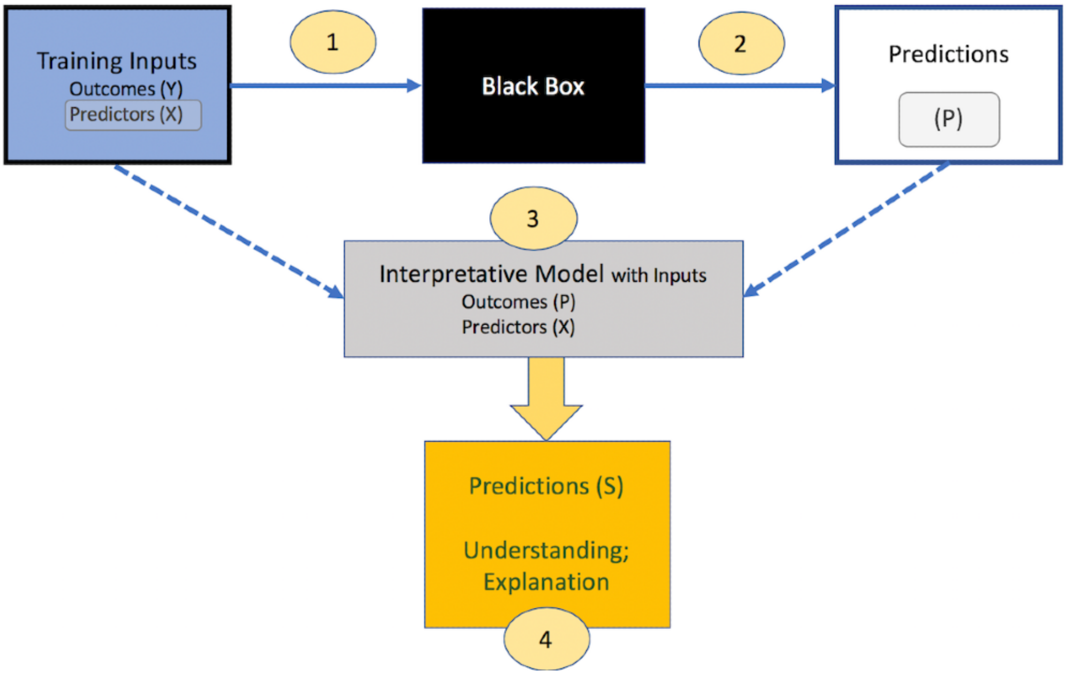
[t] LV: left ventricular.

[u] LGE: late gadolinium enhancement.

Interpretability

To communicate the results from ML models, such as our RF for SCD predictions, we develop representative interpretable summaries. As illustrated in Figure 1, the following general steps can be employed to create simplified representations of any *black box* prediction:

Figure 1. Steps to present machine learning (ML) predictions in an interpretable manner: The black box algorithm is applied to input data comprising outcomes (Y) and predictors (X) to obtain black-box predictions (P) of the input outcomes. The original X variables and the black-box predictions (P) are inputs to a simple model or algorithm, for example a single tree, whose predictions (S) are sufficiently close to (P) but more easily understood and explained.



1. Train the ML model with the input features (*X*) and the outcome of interest (*Y*).
2. Obtain the predicted values (*P*) from the ML model using cross-validation, a separate test dataset, or another data-division approach to ensure that predictions are not obtained from the same dataset used to train the model.
3. Train a simple, interpretable, and clinically understood model, such as a decision tree [34] or a linear or logistic regression model [35], using the predicted values (*P*) from the ML model as the outcome of interest and the corresponding input variables (*X*) from the original training dataset.
4. Obtain the predicted values (*S*) from the interpretive model. Calculate how close *S* is to *P*, that is how well the simplified model represents the ML model, using a measure such as R^2, defined as provided in Figure 2.

Figure 2. R^2 equation where i=1 to n observations evaluated. $S^{(i)}$ denotes the prediction for the i^{th} observation using the simplified model, $P^{(i)}$ denotes the prediction for the ith observation using the ML model, and P_{avg} denotes the average prediction from the ML model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(S^{(i)} - P^{(i)})^2}{\sum_{i=1}^{n}(P^{(i)} - P_{avg})^2}$$

Note that the interpretative tree can be grown sufficiently large such that R^2 is arbitrarily close to 1. If a simple tree has a small R^2, extra caution should be exercised to avoid overinterpreting the simplified model. In contrast, if R^2 is high, the simplified model may be considered as an alternative to the actual ML model for obtaining future predictions in a simplified manner [36,37]. This model-agnostic approach to obtain a simplified summary of the ML model is shown in Figure 1.

By using a single tree as a summary of the RF predictions, we can quantify the importance of individual variables or groups of variables. A useful measure of the total effect on outcome *Y* of predictor (or group of predictors) X_1 is obtained by summing the improvements in prediction error (deviance) over all of the X_1 splits in the interpretative tree.

To present results in other ways familiar to clinicians, predictor effects can be communicated in plots where risk ratios are presented [38]. We created plots based on the relationship between the predictor variables and predicted risks. For categorical variables, risk ratios are calculated by comparing risks for different levels of the categorical variable (eg, risk ratio=[average predicted risk for males]/[average predicted risk for females]). For continuous variables, risk ratios are calculated by comparing risks for different ranges of the continuous variable (eg, risk ratio=[average predicted risk for upper tertile of age]/[average predicted risk for lower tertile of age]). CIs for these risk ratios were generated through nonparametric bootstrap approaches [39]. All analyses were conducted using R 3.5.1 (R Foundation) [40].

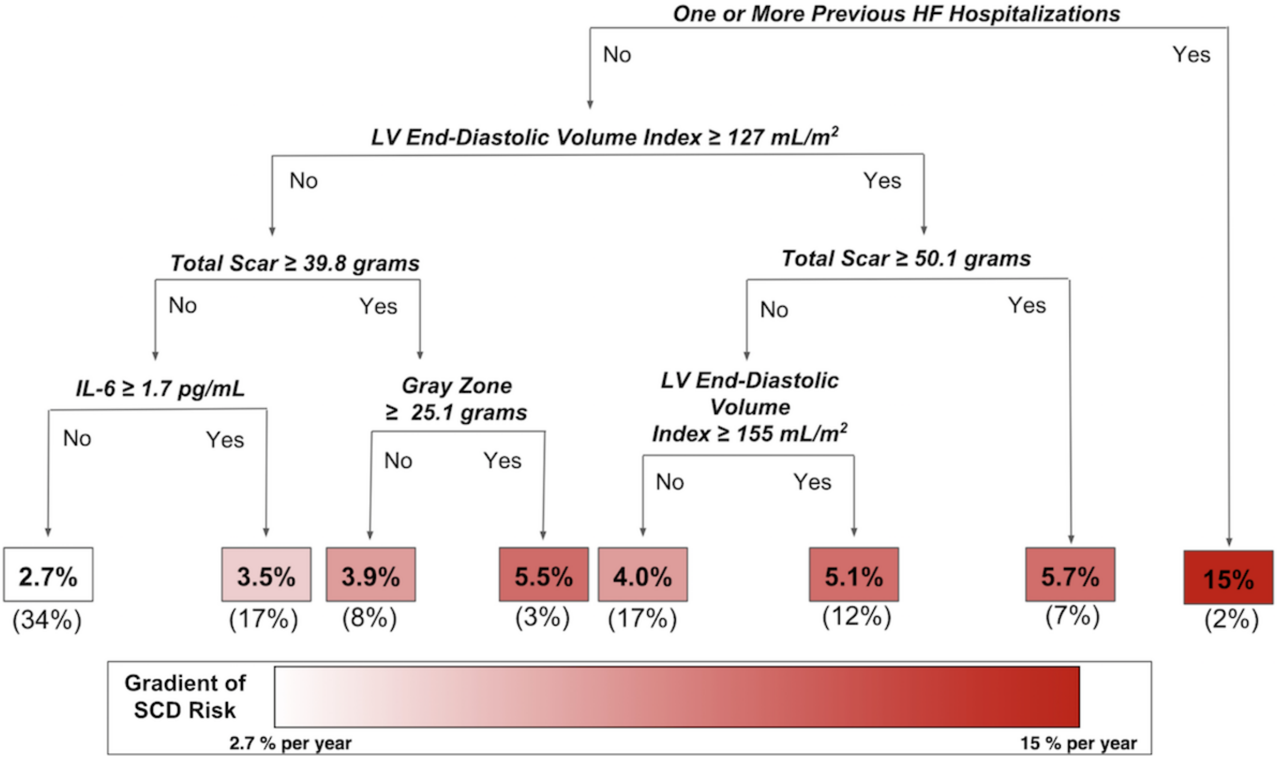
Results

Global Summary Visualization

Using data from the LV Structural Predictors of SCD Registry, a global summary for SCD risk prediction is obtained by fitting a single decision tree to RF predictions using as inputs the same covariates used in the RF and the outcome as the RF predictions.

Figure 3 shows a global summary tree of the RF model for SCD prediction. Several risk factors appear as early split nodes in the decision tree representing key variables that discriminate between low -, intermediate -, and high-risk patients, including heart failure (HF) hospitalization history, CMR imaging indices (ie, LV end-diastolic volume index, and total scar and gray zone mass), and a measure of systemic inflammation, interleukin-6 (IL-6).

Figure 3. Global summary tree of random forest (RF) model for sudden cardiac death (SCD) prediction: Several risk factors (namely heart failure hospitalization, several cardiac magnetic resonance imaging indices, and interleukin-6 [IL-6], a marker of inflammation) discriminate between low-, intermediate-, and high- risk patients. Decision rules in the tree are shown in bold italics. The 1-year risks of SCD are shown in the boxes at the bottom of the decision tree. The boxes are colored according to the magnitude of the percent per year risk, with white corresponding to the lowest risk subgroup and dark red corresponding to the highest risk subgroup. Percentages in parentheses at the bottom of the boxes are the proportions of the total training data that belong to each of the risk subgroups. R^2 is 0.88 for how well this global summary tree represents the RF model. HF: heart failure; LV: left ventricular.



Risk Ratio and Variable Importance

Figure 4 shows the risk ratio plot for predictors identified as splitting variables in the global tree summary model for our RF SCD prediction example presented in Figure 3. The largest risk ratio is for HF hospitalization history before an arrhythmic event, indicating that individuals with 1 or more preceding HF hospitalizations are at 4.06 (95% CI 2.82-5.30) times higher risk of SCD than individuals without hospitalizations for HF. Comparing the risk for individuals in the upper versus lower tertile for CMR imaging variables and IL-6 demonstrates that higher values for these variables suggest a higher SCD risk. Specifically, the risk ratios were 1.54 (95% CI 1.14-1.93) for an LV end-diastolic volume index above 133 mL/m^2 versus below 102 mL/m^2; 1.48 (95% CI 1.04-1.92) for a total scar mass above 30.79 g versus below 1.48 g; 1.48 (95% CI 1.04-1.91) for a gray zone mass above 11.37 g versus below 0.40 g, and 1.38 (95% CI 1.11-1.66) for IL-6 above 2.15 pg/mL versus below 1.04 pg/mL.

Table 2 lists the predictor variables in their order of importance in the single interpretative tree shown in Figure 3. Their ranking is based on the fraction of total variation (deviance) in the ML predictions they explain in 1 or more splits in the single tree shown in Figure 3. Although there are only 8 terminal nodes in the tree, the tree explains 88% of the information in the predictions from the *black box* RF. Additionally, trees inherently identify interactions. Note that after the first split on whether or not a person had a prior hospitalization for HF, the imaging variable only predicted risk among persons without prior hospitalization. This asymmetry indicates that the absence of a prior HF hospitalization strongly interacts with the cardiac imaging variables.

Figure 4. Visualization of predictor effects in random forest (RF) model for sudden cardiac death (SCD) prediction: Risk ratio point estimates and the 95% confidence intervals generated from 500 bootstrap replications are shown for the RF model for SCD risk prediction. The largest risk ratio is between individuals who never experienced a heart failure hospitalization and those who experienced one or more heart failure hospitalizations. The other risk ratio comparisons show the risk ratios between individuals grouped into different categories based upon inflammation or cardiac magnetic resonance (CMR) imaging variables indicating the structural and functional properties of the heart. HF: heart failure; IL-6: interleukin 6; LV: left ventricular.

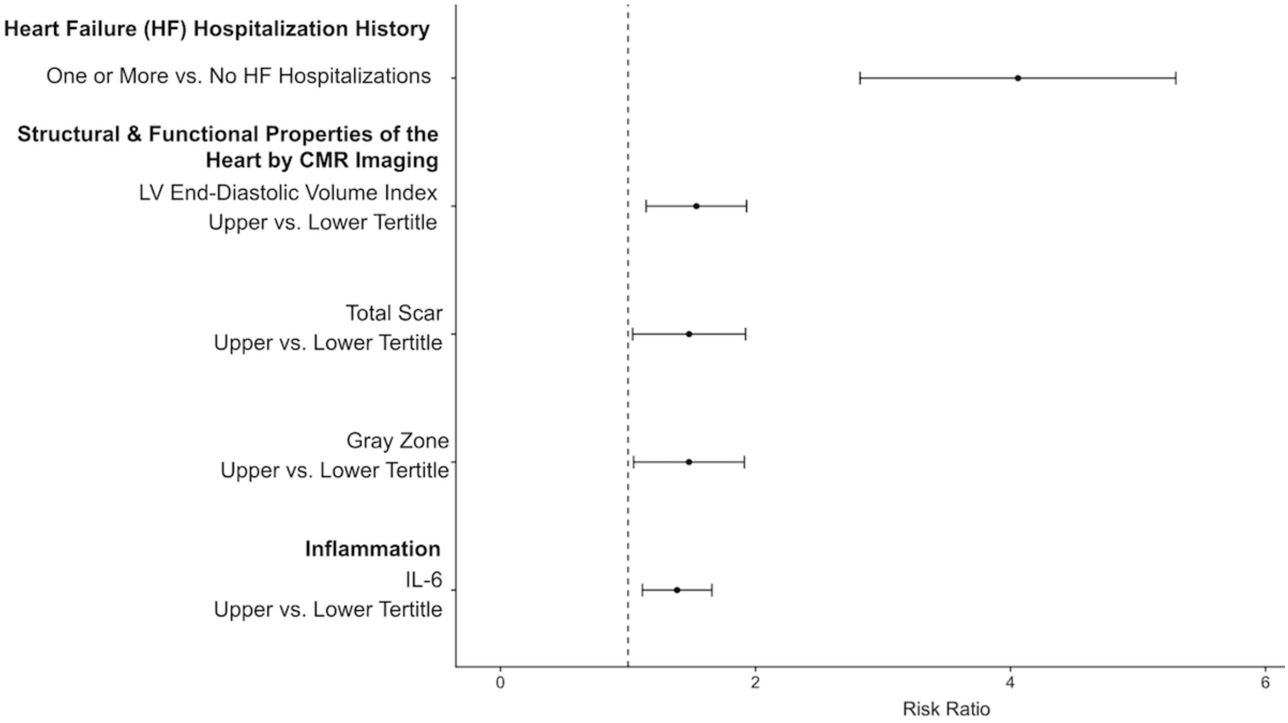


Table 2. This table summarizes the global summary tree (shown in Figure 3) with an analysis of the variation (deviance) in the predicted values (P) from the machine learning (ML) model explained by the predictors in the global summary tree. The number of splits contributed by each variable in the global summary tree is enumerated along with the deviance and the percentage of the deviance explained. The predictors' ranked importance (ordered from most to least important from left to right in the table) is determined from the percentage of the deviance explained.

Split variable	HF[a] hospitalization history	LV[b] end-diastolic volume index	Total scar	Inflammation (IL-6[c])	Gray zone	Tree total	ML[d] total
Number of splits	1	2	2	1	1	7	N/A[e]
Deviance explained	1.26	0.255	0.100	0.034	0.020	1.67	1.89
Percentage of deviance explained	66.6	13.5	5.2	1.7	1.1	0.88[f]	100

[a] HF: heart failure.

[b] LV: left ventricular.

[c] IL-6: interleukin 6.

[d] ML: machine learning.

[e] N/A: not applicable.

[f] This corresponds to the R^2 value (0.88) obtained when using the equation shown in Figure 3 for the calculations.

Discussion

Principal Findings

We demonstrate that it is possible to obtain improved transparency of ML by generating simplified models and visual displays adapted from those used commonly in clinical practice. As a specific example of this framework, we use RF extended to survival analysis with time-varying covariates for individualized SCD risk prediction. Commonly used methods for SCD risk prediction, such as Cox proportional hazards regression, do not automatically account for nonlinear and interaction effects or facilitate the application to individualized risk prediction [41,42]. In contrast to traditional regression strategies or parametric approaches that make assumptions about the underlying model, ML, such as RF, employs nonparametric algorithms that allow the data to *speak for themselves* and perform as powerful methods for individualized predictions [32,43-45]. RFs, as ensembles of decision trees, are not easily interpretable even though single decision trees are popular in medicine because of their intuitiveness and comparability to how a clinician tends to think through a case. The framework introduced in this work provides a methodology to increase ML transparency through representative interpretable summaries.

Because this framework for interpretability is model-agnostic, the user may benefit from ML's high predictive performance while also gaining insights into how predictions were generated. Despite the complexity of the original algorithm, these methods for interpretability only depend on the *inputs* upon which the *black box* was trained and the corresponding *outputs* from the *black box*, namely its predictions. Thus, any method for prediction can be explained to an extent in a simplified manner. In a situation where it is not possible to capture the variation in ML predictions with a simple summary, the proposed method signals this problem through a natural comparison of the similarity between the predicted values from the ML and its approximating interpretative model. This approach extends prior research in mimic learning and *post hoc* explanations of the *black box* predictions [46,47]. This paper emphasizes the clinician's perspective as the end user experienced with tree-based reasoning as a natural correlate of clinical reasoning.

To implement ML in clinical practice, it is essential to provide *user-centric* tools that allow clinicians to gain understanding and trust in their predictions [48]. Developing visualizations that are easy to interpret and based upon familiar ways clinicians understand algorithms or results can help communicate ML predictions. For example, simplifying RFs into a single decision tree produces a visualization that reflects medical treatment or diagnostic decision making in clinical practice. Although we illustrate the simplified model with a decision tree, other models such as linear regression can also be presented. Additionally, providing visual displays of risk ratio estimates in a manner similar to those presented in the medical literature may help clinicians gain an understanding of ML predictions.

Developing interpretable predictions is particularly important in the application of ML to health care because of the unique challenges related to medical ethics and regulatory or legal considerations [48]. Explanations that describe predictions can facilitate trust, especially when the explanations are consistent with domain knowledge or extend upon what is currently known [48]. For instance, in our illustrative example of SCD prediction in the LV Structural Predictors of SCD Registry, the key risk factors are HF hospitalization history, CMR imaging indices (ie, LV end-diastolic volume index, total scar, and gray zone mass), and a measure of inflammation (ie, IL-6). The predictors identified in our simplified summary are consistent with the published literature on SCD. It is known that among HF patients, SCD is a major cause of death due to complex interactions between the underlying myocardial substrate and triggers such as inflammation [22,49,50]. CMR imaging indices have been independently associated with ventricular arrhythmias in multiple cohorts [22,51-54]. This study raises the interaction hypothesis that cardiac imaging predictors are mainly useful in patients without prior HF hospitalizations. Visually seeing that predictions are grounded upon decision rules coinciding with clinical and biomedical knowledge (Figure 3) can help translate ML predictions for the end user's understanding. Furthermore, presenting a summary visualization of the ML model along with information about the effect estimates of the predictors (Figure 4) can facilitate further insight.

Limitations and Comparison With Prior Work

Although any complex model can be simplified to a summary model, it is possible that the summary and original model predictions are highly dissimilar, as reflected in a small R^2. This was not the case in the motivating study, where 5 variables and 7 splits explained 88% of the variation in the RF's predicted values. We can expect similar results in many problems because the interpretive tree is trained on the predicted values from a complex ML algorithm designed to find relatively lower-dimensional summaries than the original data. When a small interpretative tree has a poor R^2, it can be enlarged as needed to achieve a prespecified higher value. The user can then look for simpler summaries by grouping classes of predictors and interactions among them. Finally, the approach has the R^2 value as a measure of the fidelity of the simpler model predictions to the ML predictions. When this value is too small for a given tree, the user knows that a simple tree has limited interpretative value.

A closely related subfield of ML is actively addressing this topic by comparing different learning algorithms and selecting a final model [55]. Additionally, as the general approach for obtaining a simplified model summarizes the complex model at a global level, the simplified model is considered a global surrogate model and may not be representative of certain subgroups (eg, different subpopulations may exhibit different relationships between the predictor variables and predictions) [37]. To address this possibility, multiple simplified models could be created for each subgroup of interest. For example, two different summary decision trees could be created for men and women. Another area of active research is the development of local explanation models, where the interpretable models are local surrogate models that explain individual ML predictions rather than the entire *black box* as a whole [56]. Furthermore, although we emphasized here the tailoring of visual displays to clinicians, research in focus groups with both clinicians and patients can further accelerate the progress toward clinically meaningful ML developments that are translated into patient care.

Conclusions

Currently, limited interpretability remains a major barrier to successful translation of ML predictions to the clinical domain [1-5]. Although numerous tools such as those for feature importance, feature effect, and prediction explanations have previously been developed to facilitate interpretability [9-12,56,57], the clinical community as a whole still generally considers ML as a field that generates *black box* predictions [1-5]. Although further research is necessary to fully understand the challenges limiting the clinical implementation of these tools, we believe that emphasis on tailoring explanations and visual displays to the end user is essential. Here, we expand upon the toolkit for opening the *black box* to the clinical community through the presentation of clinically relevant and interpretable visualizations to aid in the progress toward incorporating ML in health care. Ultimately, multidisciplinary teams with combined clinical and data science expertise are essential in furthering research to address the challenges limiting the clinical implementation of these powerful, informative ML tools.

Acknowledgments

This work was supported by the National Institutes of Health Grant Nos. R01HL103812, F30HL142131, and 5T32GM007309. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

SW, KW, and SZ conceived and formulated the study design. SW and SZ developed the methods and performed the data analysis. KW acquired the patient data for the study and provided input regarding the analytic approach. SW drafted the manuscript. SW, KW, and SZ contributed to critical revision of the manuscript and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Maddox TM, Rumsfeld JS, Payne PR. Questions for Artificial Intelligence in health care. J Am Med Assoc 2019 Jan 1;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
2. Beam AL, Kohane IS. Big Data and Machine Learning in health care. J Am Med Assoc 2018 Apr 3;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
3. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of Machine Learning in medicine. J Am Med Assoc 2017 Aug 8;318(6):517-518. [doi: 10.1001/jama.2017.7797] [Medline: 28727867]
4. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the era of Artificial Intelligence. J Am Med Assoc 2018 Dec 4;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
5. Wang F, Casalino LP, Khullar D. Deep Learning in medicine-promise, progress, and challenges. JAMA Intern Med 2019 Mar 1;179(3):293-294. [doi: 10.1001/jamainternmed.2018.7117] [Medline: 30556825]
6. Lipton Z. The doctor just won't accept that!. arXiv preprint 2017 Nov 24 preprint; arXiv:1711.08037v2 [FREE Full text]
7. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable Machine Learning. arXiv preprint 2017 Mar 2 preprint; arXiv:1702.08608v2 [FREE Full text] [doi: 10.1201/9780367816377-16]
8. Adadi A, Berrada M. Peeking inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/ACCESS.2018.2870052]
9. Greenwell BM, Boehmke BC, McCarthy AJ. A simple and effective model-based variable importance measure. arXiv preprint 2018 May 12 Preprint; arXiv:1805.04755 [FREE Full text]
10. Ishwaran H. Variable importance in binary regression trees and forests. Electron J Statist 2007;1:519-537. [doi: 10.1214/07-ejs039]
11. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the Black Box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 2015;24(1):44-65. [doi: 10.1080/10618600.2014.907095]
12. Apley DW, Zhu J. Visualizing the effects of predictor variables in Black Box supervised learning models. arXiv preprint 2019 Aug 19 preprint; arXiv:1612.08468v2 [FREE Full text]
13. Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B. Rule extraction from support vector machines: an overview of issues and application in credit scoring. In: Diederich J, editor. Rule Extraction From Support Vector Machines. Berlin: Springer; 2008.
14. Stiglic G, Mertik M, Podgorelec V, Kokol P. Using Visual Interpretation of Small Ensembles in Microarray Analysis. In: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems. 2006 Presented at: CBMS'06; June 22-23, 2006; Salt Lake City, UT, USA. [doi: 10.1109/CBMS.2006.169]
15. Wolpert DH. Stacked generalization. Neural Netw 1992;5(2):241-259. [doi: 10.1016/s0893-6080(05)80023-1]
16. Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res 1999 May 1;10:271-289. [doi: 10.1613/jair.594]
17. Fishman GI, Chugh SS, Dimarco JP, Albert CM, Anderson ME, Bonow RO, et al. Sudden cardiac death prediction and prevention: report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society Workshop. Circulation 2010 Nov 30;122(22):2335-2348 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.110.976092] [Medline: 21147730]
18. Hayashi M, Shimizu W, Albert CM. The spectrum of epidemiology underlying sudden cardiac death. Circ Res 2015 Jun 5;116(12):1887-1906 [FREE Full text] [doi: 10.1161/CIRCRESAHA.116.304521] [Medline: 26044246]
19. Wellens HJ, Schwartz PJ, Lindemans FW, Buxton AE, Goldberger JJ, Hohnloser SH, et al. Risk stratification for sudden cardiac death: current status and challenges for the future. Eur Heart J 2014 Jul 1;35(25):1642-1651 [FREE Full text] [doi: 10.1093/eurheartj/ehu176] [Medline: 24801071]
20. Kandala J, Oommen C, Kern KB. Sudden cardiac death. Br Med Bull 2017 Jun 1;122(1):5-15. [doi: 10.1093/bmb/ldx011] [Medline: 28444125]

21. Myerburg RJ, Goldberger JJ. Sudden cardiac arrest risk assessment: population science and the individual risk mandate. JAMA Cardiol 2017 Jun 1;2(6):689-694. [doi: 10.1001/jamacardio.2017.0266] [Medline: 28329250]
22. Wu KC. Sudden cardiac death substrate imaged by magnetic resonance imaging: from investigational tool to clinical applications. Circ Cardiovasc Imaging 2017 Jul;10(7) [FREE Full text] [doi: 10.1161/CIRCIMAGING.116.005461] [Medline: 28637807]
23. Wu KC, Gerstenblith G, Guallar E, Marine JE, Dalal D, Cheng A, et al. Combined cardiac magnetic resonance imaging and C-reactive protein levels identify a cohort at low risk for defibrillator firings and death. Circ Cardiovasc Imaging 2012 Mar;5(2):178-186 [FREE Full text] [doi: 10.1161/CIRCIMAGING.111.968024] [Medline: 22267750]
24. Schmidt A, Azevedo CF, Cheng A, Gupta SN, Bluemke DA, Foo TK, et al. Infarct tissue heterogeneity by magnetic resonance imaging identifies enhanced cardiac arrhythmia susceptibility in patients with left ventricular dysfunction. Circulation 2007 Apr 17;115(15):2006-2014 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.106.653568] [Medline: 17389270]
25. Tao S, Ashikaga H, Ciuffo LA, Yoneyama K, Lima JA, Frank TF, et al. Impaired left atrial function predicts inappropriate shocks in primary prevention implantable cardioverter-defibrillator candidates. J Cardiovasc Electrophysiol 2017 Jul;28(7):796-805 [FREE Full text] [doi: 10.1111/jce.13234] [Medline: 28429529]
26. Zhang Y, Guallar E, Weiss RG, Stillabower M, Gerstenblith G, Tomaselli GF, et al. Associations between scar characteristics by cardiac magnetic resonance and changes in left ventricular ejection fraction in primary prevention defibrillator recipients. Heart Rhythm 2016 Aug;13(8):1661-1666 [FREE Full text] [doi: 10.1016/j.hrthm.2016.04.013] [Medline: 27108939]
27. Cheng A, Dalal D, Butcher B, Norgard S, Zhang Y, Dickfeld T, et al. Prospective observational study of implantable cardioverter-defibrillators in primary prevention of sudden cardiac death: study design and cohort description. J Am Heart Assoc 2013 Feb 22;2(1):e000083 [FREE Full text] [doi: 10.1161/JAHA.112.000083] [Medline: 23525420]
28. Cheng A, Zhang Y, Blasco-Colmenares E, Dalal D, Butcher B, Norgard S, et al. Protein biomarkers identify patients unlikely to benefit from primary prevention implantable cardioverter defibrillators: findings from the Prospective Observational Study of Implantable Cardioverter Defibrillators (PROSE-ICD). Circ Arrhythm Electrophysiol 2014 Dec;7(6):1084-1091 [FREE Full text] [doi: 10.1161/CIRCEP.113.001705] [Medline: 25273351]
29. Zhang Y, Guallar E, Blasco-Colmenares E, Dalal D, Butcher B, Norgard S, et al. Clinical and serum-based markers are associated with death within 1 year of de novo implant in primary prevention ICD recipients. Heart Rhythm 2015 Feb;12(2):360-366 [FREE Full text] [doi: 10.1016/j.hrthm.2014.10.034] [Medline: 25446153]
30. Ishwaran H, Kogalur UB. The Comprehensive R Archive Network. 2020. randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC) URL: https://cran.r-project.org/package=randomForestSRC [accessed 2017-03-10]
31. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. BMC Med Res Methodol 2019 Dec 31;20(1):1 [FREE Full text] [doi: 10.1186/s12874-019-0863-0] [Medline: 31888507]
32. Fernández-Delgado M, Cernadas E, Barro S, Amorim D, Fernández-Delgado A. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014 Jan;15(1) [FREE Full text]
33. Breiman L. Random forests. Mach Learn 2001 Oct;45:5-32. [doi: 10.1023/A%3A1010933404324]
34. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Boca Raton, FL: Wadsworth International Group; 1993.
35. Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.
36. Molnar C. Interpretable Machine Learning. San Francisco: GitHub; 2019.
37. Hall P, Phan W, Ambati S. O'Reilly Media. Boston: O'Reilly; 2017 Mar 15. Ideas on Interpreting Machine Learning URL: https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/ [accessed 2020-03-23]
38. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. Br Med J 2001 Jun 16;322(7300):1479-1480 [FREE Full text] [doi: 10.1136/bmj.322.7300.1479] [Medline: 11408310]
39. Efron B. An Introduction To The Bootstrap. New York: Chapman & Hall Crc Press; 1994.
40. R Foundation. The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed 2020-03-23]
41. Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Statist Surv 2011;5:44-71. [doi: 10.1214/09-ss047]
42. Yu C, Greiner R, Lin H, Baracos V. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011 Presented at: NIPS'11; December 16 - 17, 2011; Granada, Spain.
43. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814 [FREE Full text] [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]
44. Bzdok D. Classical statistics and statistical learning in imaging neuroscience. Front Neurosci 2017;11:543 [FREE Full text] [doi: 10.3389/fnins.2017.00543] [Medline: 29056896]
45. Bzdok D, Krzywinski M, Altman N. Points of Significance: Machine learning: a primer. Nat Methods 2017 Nov 30;14(12):1119-1120 [FREE Full text] [doi: 10.1038/nmeth.4526] [Medline: 29664466]

46. Ribeiro M, Singh S, Guestrin C. Model-agnostic interpretability of Machine Learning. arXiv preprint 2016 Jun 16 preprint; arXiv:1606.05386v1 [FREE Full text]
47. Du M, Liu N, Hu X. Techniques for interpretable machine learning. Commun ACM 2019 Dec;63(1):68-77. [doi: 10.1145/3359786]
48. Ahmad MA, Teredesai A, Eckert C. Interpretable Machine Learning in Healthcare. In: Proceedings of the 2018 IEEE International Conference on Healthcare Informatics. 2018 Presented at: ICHI'18; June 4-7, 2018; New York, NY, USA. [doi: 10.1109/ICHI.2018.00095]
49. Hussein AA, Gottdiener JS, Bartz TM, Sotoodehnia N, DeFilippi C, See V, et al. Inflammation and sudden cardiac death in a community-based population of older adults: the Cardiovascular Health Study. Heart Rhythm 2013 Oct;10(10):1425-1432. [doi: 10.1016/j.hrthm.2013.07.004] [Medline: 23906927]
50. Steinberg BA, Mulpuru SK, Fang JC, Gersh BJ. Sudden death mechanisms in nonischemic cardiomyopathies: Insights gleaned from clinical implantable cardioverter-defibrillator trials. Heart Rhythm 2017 Dec;14(12):1839-1848. [doi: 10.1016/j.hrthm.2017.09.025] [Medline: 28919378]
51. Disertori M, Rigoni M, Pace N, Casolo G, Masè M, Gonzini L, et al. Myocardial Fibrosis Assessment by LGE is a powerful predictor of Ventricular Tachyarrhythmias in ischemic and nonischemic LV dysfunction: a meta-analysis. JACC Cardiovasc Imaging 2016 Sep;9(9):1046-1055 [FREE Full text] [doi: 10.1016/j.jcmg.2016.01.033] [Medline: 27450871]
52. Jablonowski R, Chaudhry U, van der Pals J, Engblom H, Arheden H, Heiberg E, et al. Cardiovascular magnetic resonance to predict appropriate implantable cardioverter defibrillator therapy in ischemic and nonischemic cardiomyopathy patients using late gadolinium enhancement border zone: comparison of four analysis methods. Circ Cardiovasc Imaging 2017 Sep;10(9) [FREE Full text] [doi: 10.1161/CIRCIMAGING.116.006105] [Medline: 28838961]
53. Scott PA, Rosengarten JA, Curzen NP, Morgan JM. Late gadolinium enhancement cardiac magnetic resonance imaging for the prediction of ventricular tachyarrhythmic events: a meta-analysis. Eur J Heart Fail 2013 Sep;15(9):1019-1027 [FREE Full text] [doi: 10.1093/eurjhf/hft053] [Medline: 23558217]
54. Rayatzadeh H, Tan A, Chan RH, Patel SJ, Hauser TH, Ngo L, et al. Scar heterogeneity on cardiovascular magnetic resonance as a predictor of appropriate implantable cardioverter defibrillator therapy. J Cardiovasc Magn Reson 2013 Apr 10;15:31 [FREE Full text] [doi: 10.1186/1532-429X-15-31] [Medline: 23574733]
55. Bouckaert RR, Frank E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2004 Presented at: PAKDD'04; May 26-28, 2004; Sydney, Australia p. 3-12. [doi: 10.1007/978-3-540-24775-3_3]
56. Ribeiro MT, Singh S, Guestrin C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, New York, USA: ACM Press; 2016 Aug Presented at: KDD'16; August 13 - 17, 2016; San Francisco, CA p. 1135-1144.
57. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 2014;41(3):647-665. [doi: 10.1007/s10115-013-0679-x]

Abbreviations

CMR: cardiac magnetic resonance
HF: heart failure
IL-6: interleukin-6
LV: left ventricular
ML: machine learning
RF: random forest
SCD: sudden cardiac death

Edited by C Lovis; submitted 07.08.19; peer-reviewed by T Maddox, G Stiglic, D Carvalho; comments to author 16.10.19; revised version received 10.12.19; accepted 01.02.20; published 09.06.20.

Please cite as:
Wongvibulsin S, Wu KC, Zeger SL
Improving Clinical Translation of Machine Learning Approaches Through Clinician-Tailored Visual Displays of Black Box Algorithms: Development and Validation
JMIR Med Inform 2020;8(6):e15791
URL: https://medinform.jmir.org/2020/6/e15791
doi:10.2196/15791
PMID:32515746

Original Paper

Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development

Can Hou[1*], MPH, DPhil; Xiaorong Zhong[2,3*], DPhil, MD; Ping He[2,3], DPhil, MD; Bin Xu[1], MSc; Sha Diao[1], MSc; Fang Yi[1], MSc; Hong Zheng[2,3*], DPhil, MD; Jiayuan Li[1*], DPhil

[1]Department of Epidemiology and Biostatistics, West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

[2]Department of Head, Neck and Mammary Gland Oncology, Cancer Center, West China Hospital, Sichuan University, Chengdu, China

[3]Laboratory of Molecular Diagnosis of Cancer, Clinical Research Center for Breast, West China Hospital, Sichuan University, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Jiayuan Li, DPhil
Department of Epidemiology and Biostatistics
West China School of Public Health and West China Fourth Hospital
Sichuan University
No.16 Ren Min Nan Lu
Chengdu
China
Phone: 86 189 8060 1830
Email: lijiayuan73@163.com

Abstract

Background: Risk-based breast cancer screening is a cost-effective intervention for controlling breast cancer in China, but the successful implementation of such intervention requires an accurate breast cancer prediction model for Chinese women.

Objective: This study aimed to evaluate and compare the performance of four machine learning algorithms on predicting breast cancer among Chinese women using 10 breast cancer risk factors.

Methods: A dataset consisting of 7127 breast cancer cases and 7127 matched healthy controls was used for model training and testing. We used repeated 5-fold cross-validation and calculated AUC, sensitivity, specificity, and accuracy as the measures of the model performance.

Results: The three novel machine-learning algorithms (XGBoost, Random Forest and Deep Neural Network) all achieved significantly higher area under the receiver operating characteristic curves (AUCs), sensitivity, and accuracy than logistic regression. Among the three novel machine learning algorithms, XGBoost (AUC 0.742) outperformed deep neural network (AUC 0.728) and random forest (AUC 0.728). Main residence, number of live births, menopause status, age, and age at first birth were considered as top-ranked variables in the three novel machine learning algorithms.

Conclusions: The novel machine learning algorithms, especially XGBoost, can be used to develop breast cancer prediction models to help identify women at high risk for breast cancer in developing countries.

(JMIR Med Inform 2020;8(6):e17364) doi:10.2196/17364

KEYWORDS

machine learning; XGBoost; random forest; deep neural network; breast cancer

Introduction

In China, female breast cancer is the most prevalent malignant tumor affecting women, and its incidence is still increasing. According to the National Central Cancer Registry of China, more than 279,000 women were diagnosed with breast cancer in 2014, with a corresponding age-adjusted incidence rate of 28.77 per 100,000 [1]. The large number of breast cancer cases in China has resulted in a tremendous disease burden. In 2016, there were over 2 million disability-adjusted life years (DALYs) and 70,000 deaths due to breast cancer in China, accounting for approximately 15% of global DALYs and 13% of global deaths due to breast cancer [2]. Therefore, breast cancer is a major public health issue in China.

Breast cancer screening has proven to be an effective approach for breast cancer control. Several randomized controlled trials have shown that breast cancer screening can help detect breast cancer at an early stage and improve disease outcomes [3,4]. In many developed countries, population-based breast cancer screening programs have been implemented for several decades and brought positive results [5]. Nevertheless, due to the relatively low incidence rate, large population, and limited medical resources, population-based breast cancer screening is not feasible in China [6]. Therefore, some researchers have proposed risk-based breast cancer screening, considered to be cost-effective and more suitable for low- and middle-income countries like China [7].

Successful implementation of risk-based breast cancer screening largely relies on a breast cancer prediction model to accurately identify high-risk people before screening, but there is currently no suitable breast cancer prediction model for Chinese women. Some well known and commonly used breast cancer prediction models like the Gail and Tyrer-Cuzick models were developed based on women living in western countries, and their performance in Chinese women is unsatisfactory [8]. Hence, there is an urgent need to develop a breast cancer prediction model specifically for Chinese women.

Despite conventional statistical methods and some traditional machine learning algorithms (eg, logistic regression [LR]), modern machine learning has become an alternative approach for developing prediction models. Different from traditional prediction models where relationships between dependent and independent variables are predefined using prior knowledge, modern machine learning can automatically learn the underlying patterns of the data without any implicit assumptions [9]. This is especially the case for tree-based machine learning algorithms such as decision trees. These algorithms only make weak assumptions about the form of the mapping function and are therefore free to learn any functions underlying the training data and can deal with nonlinear relationships and higher order interactions between variables [10], both of which are common challenges in the health care field. In contrast, as a form of parametric machine learning algorithm, an artificial neural network (ANN) also makes strong assumptions about the functional form but it can still be used for modeling nonlinear relationships and high-order interactions. This is mainly due to the use of nonlinear activation functions in ANN and the sufficient complexity (depth and number of neurons) of the networks [11].

The objectives of this study are to evaluate and compare the performance of four different machine learning algorithms on predicting breast cancer among Chinese women and choose the best machine learning algorithm to develop a breast cancer prediction model. We used three novel machine learning algorithms in this study: extreme gradient boosting (XGBoost), random forest (RF), and deep neural network (DNN), with traditional LR as a baseline comparison.

Methods

Dataset and Study Population

In this study, we used a balanced dataset for training and testing the four machine learning algorithms. The dataset comprises 7127 breast cancer cases and 7127 matched healthy controls. Breast cancer cases were derived from the Breast Cancer Information Management System (BCIMS) at the West China Hospital of Sichuan University. The BCIMS contains 14,938 breast cancer patient records dating back to 1989 and includes information like patient characteristics, medical history, and breast cancer diagnosis [12]. West China Hospital of Sichuan University is a government-owned hospital and has the highest reputation in terms of cancer treatment in Sichuan province; the cases derived from the BCIMS are representative of breast cancer cases in Sichuan [12].

Han Chinese women living in Sichuan province who were first diagnosed with primary breast cancer between 2000 and 2017 were included (12,175 cases were included and 2763 cases were excluded). We excluded cases of patients with mental disorder and aged younger than 30 years or older than 70 years (11,916 cases were included and 259 cases were excluded). For the remaining cases, those containing missing values (4,771 cases) or contradictory data (18 cases; eg, age at first birth < age of menarche) were also excluded. Finally, a total of 7127 cases were eligible and included in the study. For each eligible breast cancer case, a main residence (urban or rural area) matched healthy control was selected from women who participated in the Breast Cancer Screening Cohort in Sichuan from 2009 to 2017. The screening project was launched at Chengdu Women’s and Children’s Central Hospital and Shuangliu Maternal and Child Health Hospital with the purpose of providing free screening for breast cancer, cervical cancer, and reproductive tract infections to women aged between 30 and 70 years.

A total of 13,607 women living in Chengdu and 15,704 women living in Shuangliu county were recruited in the cohort, representing Han Chinese women living in the developed and less developed regions of Sichuan province, respectively. Eligibility criteria for controls were Han Chinese, living in Sichuan province, confirmed to have no breast cancer, and without mental disorder and other malignant tumors. If a woman had more than one screening record, the most recent record was used.

Variable Selection

For the controls, a standard questionnaire was used to collect demographic and risk factor information, whereas for the cases, the corresponding data were directly extracted from the BCIMS. Independent variables that were included in the machine learning models were selected based on the following criteria: (1) they must be potential or known breast cancer risk factors and (2) they must be collected using the same measurement methods and have the same definitions in the cases and controls. A total of 10 variables, including 3 demographic factors, 6 reproductive history factors, and family history of breast cancer, met the above criteria and were selected for classification. Table 1 shows the details of the 10 independent variables selected along with the outcome variable.

Table 1. Descriptions and details of the variables included in the machine learning algorithms.

Variable	Types of variable	Description
Main residence	Categorical variable[a]	Urban area (0), rural area (1)
Menopausal status	Categorical variable	Premenopause (0), postmenopause (1)
Age in years	Discrete variable	Age at breast cancer diagnosis or screening
BMI (kg/m^2)	Continuous variable	BMI at breast cancer diagnosis or screening
Age of menarche	Discrete variable	Age at first menstruation
Duration of reproductive life span	Discrete variable	Premenopausal women: current age – age of menarche; postmenopausal women: menopause age – age of menarche
Pregnancy history	Categorical variable	No (0), yes (1)
Number of live births	Discrete variable	Live births is defined as births of children who showed any sign of life
Age at first birth	Discrete variable	Age of women at birth of first child (for women with no live birth, this value equals 99)
Family history of breast cancer	Categorical variable	First-degree or second-degree female relatives had breast cancer: no (0), yes (1)
Case-control status (outcome variable)	Categorical variable	Control (0), case (1)

[a]Categorical variables were converted into one-hot encoding before being provided to machine learning algorithms.

Machine Learning Algorithms

In this study, three novel machine learning algorithms (XGBoost, RF, and DNN) along with a baseline comparison (LR) were evaluated and compared.

XGBoost and RF both belongs to ensemble learning, which can be used for solving classification and regression problems. Different from ordinary machine learning approaches where only one learner is trained using a single learning algorithm, ensemble learning consists of many base learners. The predictive performance of a single base learner is merely slightly better than random guess, but ensemble learning can boost them to strong learners with high prediction accuracy by combination [13]. There are two methods to combine base learners: bagging and boosting. The former is the base of RF while the latter is the base of XGBoost. In RF, decision trees are used as base learners and bootstrap aggregating, or bagging, is used to combine them [14]. XGBoost is based on the gradient boosted decision tree (GBDT), which uses decision trees as base learners and gradient boosting as combination method. Compared with GBDT, XGBoost is more efficient and has better prediction accuracy due to its optimization in tree construction and tree searching [15].

DNN is an ANN with many hidden layers [16]. A standard ANN is made up of an input layer, several hidden layers, and an output layer, and each layer contains multiple neurons. Neurons in the input layer receive values from the input data, neurons in other layers receive weighted values from the previous layers and apply nonlinearity to the aggregation of the values [16]. The learning process is to optimize the weights using a backpropagation method to minimize the differences between predicted outcomes and true outcomes. In contrast to shallow ANN, DNN can learn more complex nonlinear relationships and is intrinsically more powerful [17].

Hyperparameters Tuning, Model Development, and Algorithm Comparison

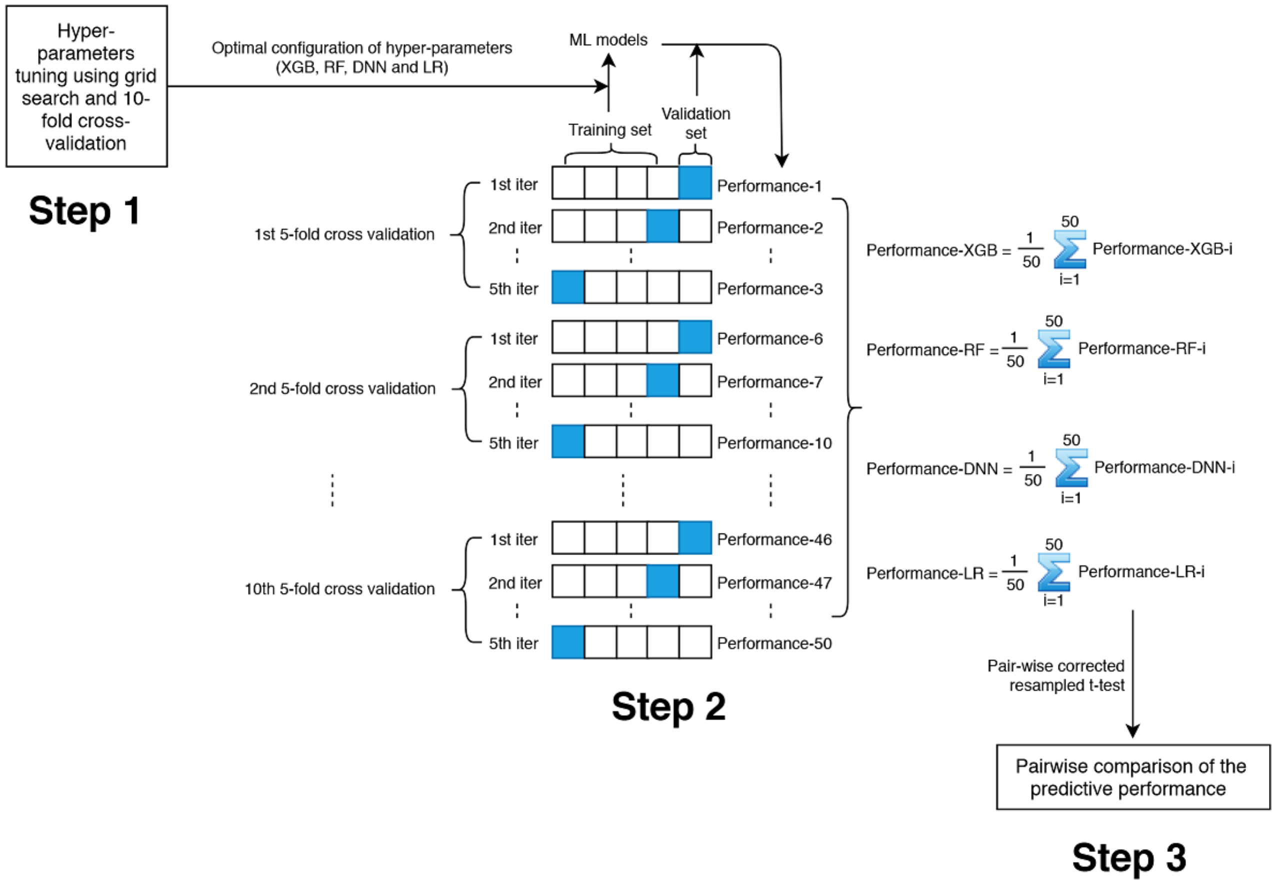
A general overview of the model development and algorithm comparison process is illustrated in Figure 1. The first step was hyperparameters tuning, with the purpose of choosing the most optimal configuration of hyperparameters for each machine learning algorithm. In DNN and XGBoost, we introduced dropout and regularization techniques, respectively, to avoid overfitting, whereas in RF, we tried to reduce overfitting by tuning the hyperparameter min_samples_leaf. We conducted a grid search and 10-fold cross-validation on the whole dataset for hyperparameters tuning. The results of the hyperparameters tuning along with the optimal configuration of hyperparameters for each machine learning algorithm is shown in Multimedia Appendix 1.

Based on the optimal configuration of hyperparameters, the next step was model development and assessment. In this step, we used repeated 5-fold cross-validation. This method can avoid overfitting and increase robustness of the results. In each 5-fold cross-validation, the dataset was randomly divided into 5 folds with approximately equal sample size, where 4 folds were chosen as training set to develop the machine learning models while the remaining 1 fold was used as the validation set to calculate the model performance metrics (including area under the receiver operating characteristic curve [AUC], sensitivity, specificity, and accuracy). After 5 iterations, each fold (as well as each subject) was used as validation set exactly once. We repeated the whole 5-fold cross-validation process 10 times, and in each repetition, the division of the dataset was different.

The final step was algorithm comparison. For each machine learning algorithm, we summarized their predictive performance metrics generated from the second step using means and standard deviations and conducted pair-wise comparison using statistical tests. AUC was chosen as the primary measure of the predictive performance in our study.

RF and LR algorithms were implemented in Python 3.6 (Python Software Foundation) using scikit-learn (version 0.20.0). XGBoost and DNN algorithms were implemented in Python 3.6 using xgboost (version 0.80) and TensorFlow (version 1.10.0), respectively. Source code is shown in Multimedia Appendix 2.

Figure 1. Process of model development and algorithm comparison. Step 1: hyperparameters tuning; step 2: model development and assessment; step 3: algorithm comparison. Performance metrics include area under the receiver operating characteristic curve, sensitivity, specificity, and accuracy.


Hyper-parameters tuning using grid search and 10-fold cross-validation
Step 1
Optimal configuration of hyper-parameters (XGB, RF, DNN and LR)
ML models
Training set
Validation set
1st 5-fold cross validation
2nd 5-fold cross validation
10th 5-fold cross validation
1st iter
2nd iter
5th iter
Performance-1
Performance-2
Performance-3
Performance-6
Performance-7
Performance-10
Performance-46
Performance-47
Performance-50
Performance-XGB = 1/50 Σ_{i=1}^{50} Performance-XGB-i
Performance-RF = 1/50 Σ_{i=1}^{50} Performance-RF-i
Performance-DNN = 1/50 Σ_{i=1}^{50} Performance-DNN-i
Performance-LR = 1/50 Σ_{i=1}^{50} Performance-LR-i
Step 2
Pair-wise corrected resampled t-test
Pairwise comparison of the predictive performance
Step 3

Variable Ranking

To have a deeper understanding of the three novel machine learning algorithms, we ranked all the variables based on their impact on AUCs. To do so, we repeated step 1 and 2 illustrated above, but in each iteration of the repeated 5-fold cross-validation, we successively permuted the values of the 10 variables in the testing set and calculated the corresponding decrease in the AUC (in percentage). Then the results were summarized and used for ranking each variable in the four machine learning algorithms. Detailed process is illustrated in Multimedia Appendix 3.

Statistical Analysis

Characteristics of the cases and controls were described using medians (IQRs) for continuous or discrete variables and number (%) for categorical variables. For the comparison of characteristics between cases and controls, we used a Mann-Whitney *U* test for continuous or discrete variables and Pearson chi-square test for categorical variables, with a significance level of .05. As for the pair-wise comparison of the predictive performance of the machine learning algorithms, we used the pair-wise corrected resampled *t* test [18]. To counteract the issue of multiple comparisons, the significance level was adjusted to .008 using Bonferroni correction. All statistical analyses were conducted using SciPy (version 1.1.0), pandas (version 0.23.0), and NumPy (version 1.14.3) in Python 3.6.

Results

As shown in Table 2, a total of 7127 breast cancer cases along with 7127 matched healthy controls were included in the dataset. Among the cases, 61.27% (4367/7127) were premenopausal women and 38.73% (2760/7127) were postmenopausal women, while among the controls, 63.80% (4547/7127) were premenopausal women and 36.20% (2580/7127) were postmenopausal women. Except for BMI, menarche age, main residence, and family history of breast cancer, all other features were significantly different between the cases and controls.

The predictive performance of the four machine learning algorithms is shown in Figure 2 and Table 3, and the results of the pair-wise corrected resampled *t* test are shown in Multimedia Appendix 4. The three novel machine learning algorithms all achieved significantly higher AUCs than the linear LR algorithm (*P*<.001). To be more specific, XGBoost had a mean score of 0.742 in terms of AUC, followed by DNN (0.728), RF (0.728), and LR (0.631). XGBoost had significantly higher AUC than both DNN and RF (*P*<.001), whereas no statistically significant difference was found between the mean AUCs of DNN and RF (*P*=.98). Similarly, the mean sensitivity of LR (49.6%) was significantly lower compared with that of XGBoost (65.6%), DNN (64.2%), or RF (65.0%; *P*<.001). However, for XGBoost, DNN, and RF, their mean sensitivities were not significantly different from each other.

As for specificity, the four machine learning algorithms achieved similar results (LR: 66.1%; DNN: 67.9%; XGBoost: 68.6%; RF: 67.7%) and no statistically significant differences were found between them. Since we used a balanced dataset to train the models, we have also reported accuracy. Compared with XGBoost (67.1%), DNN (66.1%), and RF (66.3%), LR had the lowest accuracy (57.8%). Although XGBoost had the highest mean accuracy, there was no significant difference between the mean accuracy of XGBoost and RF (*P*=.34). The difference between the mean accuracy of DNN and RF was also statistically insignificant (*P*=.01).

Figure 3 presents the variable rankings according to the mean decrease in AUCs in different machine learning algorithms. XGBoost, RF, and DNN were very similar in variable rankings, although some discrepancies did exist. In all the three novel machine learning algorithms, main residence, number of live births, menopause status, age, and age at first birth were considered as top-ranked variables. Since the cases and controls were matched by main residence, linear LR prioritized all other variables over main residence. Moreover, pregnancy history, which was not present in top-ranked variables for the three novel machine learning algorithms, was prioritized over age and age at first birth in LR.

Table 2. Characteristics of case and control participants.

Variable	Control	Case	*P* value[a]
Age in years, median (IQR)	47 (41-53)	47 (42-54)	<.001
BMI (kg/m^2), median (IQR)	22.83 (20.96-24.77)	22.89 (20.94-24.97)	.16
Age of menarche, median (IQR)	14 (13-15)	14 (13-15)	.23
Duration of reproductive lifespan, median (IQR)	31 (27-34)	32 (27-35)	<.001
Number of live births, median (IQR)	1 (1-1)	1 (1-2)	<.001
Age at first birth[b], median (IQR)	24 (22-26)	24 (23-26)	<.001
Main residence, n (%)			>.99
Urban area	3994 (56.04)	3994 (56.04)	
Rural area	3133 (43.96)	3133 (43.96)	
Pregnancy history, n (%)			<.001
No	201 (2.82)	14 (0.20)	
Yes	6926 (97.18)	7113 (99.80)	
Family history of breast cancer, n (%)			.17
No	7010 (98.36)	6988 (98.05)	
Yes	117 (1.64)	139 (1.95)	
Menopausal status, n (%)			<.01
Premenopause	4367 (61.27)	4547 (63.80)	
Postmenopause	2760 (38.73)	2580 (36.20)	

[a] *P* values are derived from Mann-Whitney *U* test or Pearson chi-square test.

[b] Only women with at least one birth are summarized.

Figure 2. Mean receiver operating characteristic curves for the four machine learning algorithms.

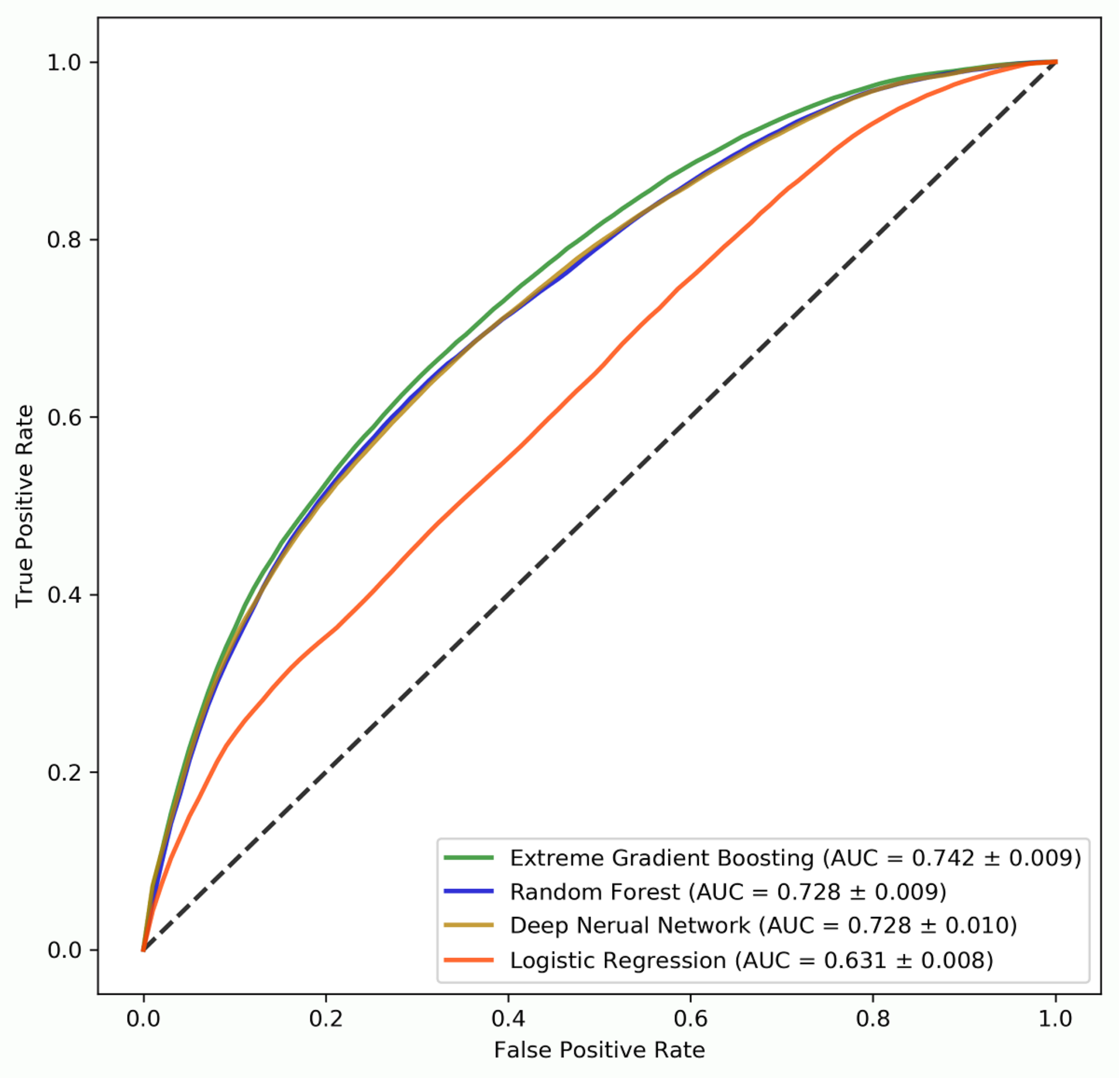


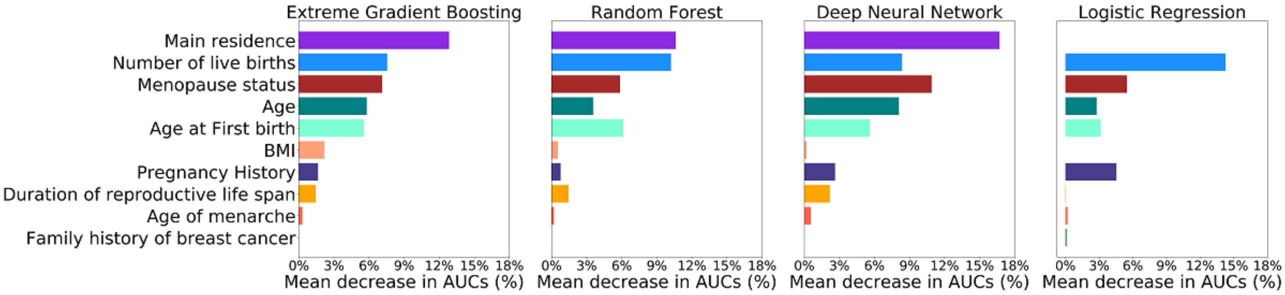
Table 3. Performance of the four machine learning algorithms on predicting breast cancer risk.

Algorithm	AUC[a], mean (SD)	Sensitivity[b], mean (SD)	Specificity[b], mean (SD)	Accuracy[b], mean (SD)
Extreme gradient boosting	0.742 (0.009)	0.656 (0.017)	0.686 (0.012)	0.671 (0.009)
Random forest	0.728 (0.009)	0.650 (0.016)	0.677 (0.015)	0.663 (0.010)
Deep neural network	0.728 (0.010)	0.642 (0.037)	0.679 (0.033)	0.661 (0.010)
Logistic regression	0.631 (0.008)	0.496 (0.020)	0.661 (0.021)	0.578 (0.008)

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

[b] Sensitivity, specificity, and accuracy were calculated using the default cutoff value (0.5).

Figure 3. Variable rankings according to the mean area under the receiver operating characteristic curve decrease in percentage in the four machine learning algorithms.



Discussion

Principal Findings

In this study, we used four machine learning algorithms to develop breast cancer prediction models to identify Chinese women at high risk of breast cancer, based on 10 breast cancer risk factors. Their predictive performances were evaluated and compared, and the results indicated that compared with traditional LR, all three novel machine learning algorithms achieved better performance and improved the AUC by 0.11 at most. Among the three novel machine learning algorithms, XGBoost outperformed RF and DNN, with mean AUC and accuracy of 0.74 and 67.1%, respectively.

Among the three novel machine learning algorithms used in this study, XGBoost is the most up to date. Recently, XGBoost has dominated many data mining competitions for structured datasets and gained much attention in the machine learning field. Some previous studies have shown that XGBoost has better performance on low-dimensional data than high-dimensional data [19,20]. RF, on the other hand, is more suitable for high-dimensional data due to its implicit feature selection characteristic [21]. As for DNN, it is more commonly used for prediction with unstructured data and data with complex structure [22]. Therefore, the results of this study were expected, since the structure of our dataset agrees with the XGBoost algorithm most.

Meanwhile, XGBoost was also faster than DNN and RF. The average times of training XGBoost, DNN, and RF using the current dataset were 0.20 seconds, 21.38 seconds, and 0.61 seconds, respectively (CPU: Intel i7-4790; GPU: GeForce GTX 970). Given the above, XGBoost is no doubt the optimal choice for developing a breast cancer prediction model using traditional breast cancer risk factors. Nevertheless, DNN and RF are also powerful machine learning algorithms and could be considered for developing a breast cancer prediction model in other circumstances. For example, if the dataset contains high-dimensional genetic data, RF is very likely to be the best choice. As for DNN, it can be used to predict breast cancer based on breast ultrasound or mammogram images when integrated with a convolutional neural network (CNN).

Although the predictive accuracy of the novel machine learning algorithms is still imperfect, some remarkable improvements have been made compared with previous breast cancer risk prediction models. The Gail model is the most well-known breast cancer risk assessment tool. It uses six breast cancer risk factors to estimate a women's risk of developing breast cancer, including patient demographics, reproductive history, personal medical history, and family history of breast cancer [23]. Among these risk factors, four risk factors (age, age of menarche, age at first birth, and family history of breast cancer) are also present in this study. A recent meta-analysis of 26 studies reported a pooled AUC of 0.59 (95% CI 0.57 to 0.61) for the Gail model, which is significantly lower than the AUCs of XGBoost, RF, and DNN algorithms in our study. The authors also conducted subgroup analyses by geographic region, and the results revealed that the pooled AUC for the Gail model in Asian women was even lower (0.55, 95% CI 0.52 to 0.58).

Another famous breast cancer risk prediction model is the Rosner-Colditz model. This model is more complex than the Gail model and includes some key risk factors omitted from the Gail model such as type of menopause, BMI, and duration and type of postmenopausal hormone therapy used [24]. A validation study conducted by Rosner et al [25] using the dataset from the California Teachers Study revealed that the Rosner-Colditz model achieved an overall AUC of 0.59, higher than the AUC of the Gail model when applied in the same dataset. Different from the machine learning models in this study, classical breast cancer risk prediction models like the Gail and Rosner-Colditz models put more of an emphasis on estimating the probability of having breast cancer in a defined age interval instead of identifying breast cancer cases from noncases. In addition, all these models are based on an implicit assumption that each risk factor has a linear relationship with breast cancer and therefore largely ignore complex nonlinear relationships between risk factors and breast cancer and interactions between risk factors.

Many studies have been conducted to evaluate the performance of machine learning algorithms for breast cancer prediction. However, the majority of these studies used medical imaging data to develop the models, and only few focused on prediction with breast cancer risk factors. Shieh et al [26] conducted a nested case-control study in the United States to investigate the predictive performance of combining the Breast Cancer Surveillance Consortium (BCSC) risk model with an 83-single nucleotide polymorphism (SNP)–based polygenic risk score (PRS) in an LR model. They reported that compared with using the BCSC model alone, the LR model combining BCSC risk factors and PRS increased the AUC from 0.62 to 0.65. Dite et al [27] also evaluated the performance of an LR model that combined several breast cancer prediction models with a risk score based on 77 SNPs. They reported that the LR model using the absolute risk from the Breast and Ovarian Analysis of

Disease Incidence and Carrier Estimation Algorithm model and the SNP-based score achieved the best performance (AUC: 0.70). Anothaisintawee et al [28] developed an LR model for predicting breast cancer risk among Thai women using the dataset from a cross-sectional study. The variables used in building the model included age, menopausal status, BMI, and use of oral contraceptives. The LR model achieved AUCs of 0.65 (95% CI 0.64 to 0.65) and 0.61 (95% CI 0.51 to 0.71) on the internal validation and external validation datasets, respectively. Zheng et al [29] also developed an LR model for predicting breast cancer. They derived the dataset from a case-control study in China and used 12 SNPs along with age at menarche, age at first live birth, waist-to-hip ratio, family history of breast cancer, and a previous diagnosis of benign breast disease for model building. The final LR model had an AUC of 0.63, which the authors believe is inadequate for cancer diagnosis and screening. Zhao et al [30] conducted the only study that evaluated the performance of a machine learning algorithm other than LR. They built an ANN model with one hidden layer for Chinese women using a cross-sectional dataset. In the test set, the ANN model achieved an AUC of 0.71 (95% CI 0.66 to 0.76). Compared with the previous machine learning models, our novel machine learning model achieved higher AUCs. More importantly, our model only requires 10 breast cancer risk factors, which can be easily collected in a cost-effective manner.

One of the major disadvantages of machine learning algorithms is that they are hard to interpret, especially for DNN. In our study, we tried to investigate the independent effects of each variable on breast cancer prediction. XGBoost, DNN, and RF all identified main residence as the most important variable in the models. This finding indicates significant interactions between main residence and other risk factors, but we cannot determine how main residence is interacting with other risk factors. A possible explanation for the interactions is the differences in lifestyle and environmental conditions in rural and urban area. A previous cross-sectional survey conducted in China also reported that there were some differences in the breast cancer risk factors between urban and rural populations [31]. Other top-ranked variables in the machine learning models are also considered to be important breast cancer risk factors in Chinese women [32,33].

It is also recognized that machine learning algorithms are vulnerable to overfitting. To address this issue, we used repeated k-fold cross-validation method to evaluate the performance of the models. Meanwhile, we also tried to reduce overfitting by choosing appropriate hyperparameters and using regularization techniques. Another limitation of using machine learning algorithms for breast cancer prediction is that absolute risk of having breast cancer cannot be estimated. However, it is not necessarily the case that machine learning models are less useful than traditional absolute breast cancer prediction models. In some circumstances (eg, risk-based breast cancer screening), the primary objective is to identify women with high risk from those with lower risk rather than informing personalized absolute risk. Therefore, in that case machine learning models with higher discriminatory accuracy would be more useful.

Strengths and Limitations

To the best of our knowledge, this is the first study applying the novel machine learning algorithms to breast cancer prediction. The strengths of our study include using a large balanced dataset for model training and conducting repeated k-fold cross-validation for model evaluation. Nevertheless, our study still has several limitations. First, since an observational study design was adopted to derive the dataset, the influence of selection bias cannot be omitted. A better choice would be deriving the dataset from a large cohort study with sufficient breast cancer cases. However, considering the incidence of breast cancer, using such dataset would raise the issue of imbalanced classes that require statistical techniques to deal with. Second, due to the limitations of the dataset, only 10 breast cancer risk factors were chosen to build the model and some important risk factors like breastfeeding and history of other breast diseases were excluded, which may have influenced the performance of the models. In addition, our machine learning models were trained on cases and noncases from Sichuan province in southwest China and therefore may not be useful for women living in other parts of China. Furthermore, although we validated our models using the cross-validation method, external validation using an independent dataset was not performed in our study.

Currently, we have developed our final breast cancer prediction model using XGBoost and implemented it in a mobile phone app. Our next step is to validate this prediction model in a real-world setting and upgrade the model by including more risk factors and potentially medical imaging data. After external validation, we also plan to conduct a large observational study to evaluate the cost-effectiveness of applying this model in risk-based breast cancer screening among Chinese women.

Conclusion

Our study has shown that all three novel machine learning algorithms achieved better discriminatory accuracy on identifying women at high risk of breast cancer than LR, and XGBoost is the best choice for developing a breast cancer prediction model using breast cancer risk factors. We have successfully developed and validated a breast cancer prediction model for Chinese women using XGBoost, but external validation is needed before implementation.

Acknowledgments

This study is supported by the Graduate Students' Research and Innovation Fund of Sichuan University (2018YJSY112), National Natural Science Foundation of China (No. 81302500), People-Benefit Project of Chengdu Science and Technology Bureau (No. 2015-HM01-00049-SF), and Key Research and Development Project of the Department of Science and Technology in Sichuan

Province (No. 2017SZ0005). The funders had no role in the design of the study; collection, analysis, and interpretation of data; or writing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Hyperparameter tuning results and optimal configuration of hyperparameters for each machine learning algorithm.
[PDF File (Adobe PDF File), 1509 KB - medinform_v8i6e17364_app1.pdf]

Multimedia Appendix 2
Python code and outputs.
[PDF File (Adobe PDF File), 338 KB - medinform_v8i6e17364_app2.pdf]

Multimedia Appendix 3
Detailed process of variable ranking.
[PDF File (Adobe PDF File), 142 KB - medinform_v8i6e17364_app3.pdf]

Multimedia Appendix 4
Results of the pairwise comparison of the model predictive performance.
[PDF File (Adobe PDF File), 112 KB - medinform_v8i6e17364_app4.pdf]

References

1. Chen W, Sun K, Zheng R, Zeng H, Zhang S, Xia C, et al. Cancer incidence and mortality in China, 2014. Chin J Cancer Res 2018 Feb;30(1):1-12 [FREE Full text] [doi: 10.21147/j.issn.1000-9604.2018.01.01] [Medline: 29545714]
2. GBD 2016 Mortality Collaborators. Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017 Sep 16;390(10100):1084-1150 [FREE Full text] [doi: 10.1016/S0140-6736(17)31833-0] [Medline: 28919115]
3. Narod S, Sun P, Wall C, Baines C, Miller A. Impact of screening mammography on mortality from breast cancer before age 60 in women 40 to 49 years of age. Curr Oncol 2014 Oct;21(5):217-221 [FREE Full text] [doi: 10.3747/co.21.2067] [Medline: 25302030]
4. Moss SM, Wale C, Smith R, Evans A, Cuckle H, Duffy SW. Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. The Lancet Oncology 2015 Sep;16(9):1123-1132. [doi: 10.1016/s1470-2045(15)00128-x] [Medline: 26206144]
5. Dowling EC, Klabunde C, Patnick J, Ballard-Barbash R, International Cancer Screening Network (ICSN). Breast and cervical cancer screening programme implementation in 16 countries. J Med Screen 2010;17(3):139-146. [doi: 10.1258/jms.2010.010033] [Medline: 20956724]
6. Wong IOL, Kuntz KM, Cowling BJ, Lam CLK, Leung GM. Cost effectiveness of mammography screening for Chinese women. Cancer 2007 Aug 15;110(4):885-895 [FREE Full text] [doi: 10.1002/cncr.22848] [Medline: 17607668]
7. Sun L, Legood R, Sadique Z, Dos-Santos-Silva I, Yang L. Cost-effectiveness of risk-based breast cancer screening programme, China. Bull World Health Organ 2018 Aug 01;96(8):568-577 [FREE Full text] [doi: 10.2471/BLT.18.207944] [Medline: 30104797]
8. Wang X, Huang Y, Li L, Dai H, Song F, Chen K. Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis. Breast Cancer Res 2018 Mar 13;20(1):18 [FREE Full text] [doi: 10.1186/s13058-018-0947-5] [Medline: 29534738]
9. Michalski R, Carbonell J, Mitchell T. Machine Learning: An Artificial Intelligence Approach. Berlin: Springer Science & Business Media; 2013.
10. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. Malaysia: Pearson Education Limited; 2016.
11. Nielsen M. Neural Networks and Deep Learning. New York: Determination Press; 2015.
12. Peng Z, Wei J, Lu X, Zheng H, Zhong X, Gao W, et al. Treatment and survival patterns of Chinese patients diagnosed with breast cancer between 2005 and 2009 in Southwest China: an observational, population-based cohort study. Medicine (Baltimore) 2016 Jun;95(25):e3865 [FREE Full text] [doi: 10.1097/MD.0000000000003865] [Medline: 27336872]
13. Zhang C, Ma Y. Ensemble Machine Learning: Methods and Applications. New York: Springer; 2012.
14. Ho T. Random decision forests: document analysis and recognition. Proc Third Int Conf Doc Analysis Recogn 1995;1. [doi: 10.1109/icdar.1995.598994]
15. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Mining 2016:785-794. [doi: 10.1145/2939672.2939785]

16. Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks 2015 Jan;61:85-117. [doi: 10.1016/j.neunet.2014.09.003]
17. Bengio Y. Learning deep architectures for AI. Foundations Trends Machine Learning 2009;2(1):1-127. [doi: 10.1561/2200000006]
18. Bouckaert R, Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. In: Dai H, Srikant R, editors. Advances in Knowledge Discovery and Data Mining. Berlin: Springer; 2004:3-12.
19. Taylor RA, Moore CL, Cheung K, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS One 2018;13(3):e0194085 [FREE Full text] [doi: 10.1371/journal.pone.0194085] [Medline: 29513742]
20. Li B, Zhang N, Wang Y, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front Genet 2018;9:237 [FREE Full text] [doi: 10.3389/fgene.2018.00237] [Medline: 30023001]
21. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 2009 Jul 10;10:213 [FREE Full text] [doi: 10.1186/1471-2105-10-213] [Medline: 19591666]
22. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. Proc 23rd Int Conf Machine learning 2006:161-168. [doi: 10.1145/1143844.1143865]
23. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999 Sep 15;91(18):1541-1548. [doi: 10.1093/jnci/91.18.1541] [Medline: 10491430]
24. Rosner B, Colditz G. Nurses' health study: log-incidence mathematical model of breast cancer incidence. J Natl Cancer Inst 1996 Mar 20;88(6):359-364. [doi: 10.1093/jnci/88.6.359] [Medline: 8609645]
25. Rosner BA, Colditz GA, Hankinson SE, Sullivan-Halley J, Lacey JV, Bernstein L. Validation of Rosner-Colditz breast cancer incidence model using an independent data set, the California Teachers Study. Breast Cancer Res Treat 2013 Nov;142(1):187-202 [FREE Full text] [doi: 10.1007/s10549-013-2719-3] [Medline: 24158759]
26. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JWT, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat 2016 Oct;159(3):513-525 [FREE Full text] [doi: 10.1007/s10549-016-3953-2] [Medline: 27565998]
27. Dite GS, MacInnis RJ, Bickerstaffe A, Dowty JG, Allman R, Apicella C, et al. Breast cancer risk prediction using clinical models and 77 independent risk-associated SNPs for women aged under 50 years: Australian Breast Cancer Family Registry. Cancer Epidemiol Biomarkers Prev 2016 Feb;25(2):359-365 [FREE Full text] [doi: 10.1158/1055-9965.EPI-15-0838] [Medline: 26677205]
28. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Srinakarin J, Woodtichartpreecha P, Hirunpat S, et al. Development and validation of a breast cancer risk prediction model for Thai women: a cross-sectional study. Asian Pac J Cancer Prev 2014;15(16):6811-6817 [FREE Full text] [doi: 10.7314/apjcp.2014.15.16.6811] [Medline: 25169530]
29. Zheng W, Wen W, Gao Y, Shyr Y, Zheng Y, Long J, et al. Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. J Natl Cancer Inst 2010 Jul 07;102(13):972-981 [FREE Full text] [doi: 10.1093/jnci/djq170] [Medline: 20484103]
30. Zhao Y, Xiong P, McCullough LE, Miller EE, Li H, Huang Y, et al. Comparison of breast cancer risk predictive models and screening strategies for Chinese women. J Womens Health (Larchmt) 2017 Mar;26(3):294-302. [doi: 10.1089/jwh.2015.5692] [Medline: 28263689]
31. Wang F, Yu L, Wang F, Liu L, Guo M, Gao D, et al. Risk factors for breast cancer in women residing in urban and rural areas of eastern China. J Int Med Res 2015 Dec;43(6):774-789. [doi: 10.1177/0300060515592901] [Medline: 26475794]
32. Yu Z, Jia C, Geng C, Tang J, Zhang J, Liu L. Risk factors related to female breast cancer in regions of Northeast China: a 1:3 matched case-control population-based study. Chin Med J (Engl) 2012 Mar;125(5):733-740. [Medline: 22490565]
33. Lee H, Li J, Fan J, Li J, Huang R, Zhang B, et al. Risk factors for breast cancer among Chinese women: a 10-year nationwide multicenter cross-sectional study. J Epidemiol 2014;24(1):67-76 [FREE Full text] [doi: 10.2188/jea.je20120217] [Medline: 24270059]

Abbreviations

ANN: artificial neural network
AUC: area under the receiver operating characteristic curve
BCIMS: Breast Cancer Information Management System
BCSC: Breast Cancer Surveillance Consortium
DALY: disability-adjusted life year
DNN: deep neural network
GBDT: gradient boosted decision tree
LR: logistic regression
PRS: polygenic risk score

RF: random forest
SNP: single nucleotide polymorphism
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 10.12.19; peer-reviewed by Y Wang, M Kebede; comments to author 10.03.20; revised version received 28.03.20; accepted 19.04.20; published 08.06.20.

Please cite as:
Hou C, Zhong X, He P, Xu B, Diao S, Yi F, Zheng H, Li J
Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development
JMIR Med Inform 2020;8(6):e17364
URL: http://medinform.jmir.org/2020/6/e17364/
doi:10.2196/17364
PMID:32510459

Original Paper

Artificial Intelligence–Based Multimodal Risk Assessment Model for Surgical Site Infection (AMRAMS): Development and Validation Study

Weijia Chen[1*], MD; Zhijun Lu[1*], MD, PhD; Lijue You[2], MS; Lingling Zhou[3], BSc; Jie Xu[4,5], MPhil; Ken Chen[1,5,6], MD, DHM

[1]Department of Anesthesiology, Rui Jin Hospital, Luwan Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]Department of Informatics, Rui Jin Hospital, Luwan Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Infection Prevention and Control, Rui Jin Hospital, Luwan Branch, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[4]VitalStrategic Research Institute, Shanghai, China

[5]Synyi Research, Shanghai, China

[6]Precision Diagnosis and Image Guided Therapy, Philips Research China, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Ken Chen, MD, DHM
Department of Anesthesiology
Rui Jin Hospital, Luwan Branch
Shanghai Jiao Tong University School of Medicine
South Chongqing Road, No 149
Shanghai
China
Phone: 86 021 63864050
Email: nutastray@gmail.com

Abstract

Background: Surgical site infection (SSI) is one of the most common types of health care–associated infections. It increases mortality, prolongs hospital length of stay, and raises health care costs. Many institutions developed risk assessment models for SSI to help surgeons preoperatively identify high-risk patients and guide clinical intervention. However, most of these models had low accuracies.

Objective: We aimed to provide a solution in the form of an Artificial intelligence–based Multimodal Risk Assessment Model for Surgical site infection (AMRAMS) for inpatients undergoing operations, using routinely collected clinical data. We internally and externally validated the discriminations of the models, which combined various machine learning and natural language processing techniques, and compared them with the National Nosocomial Infections Surveillance (NNIS) risk index.

Methods: We retrieved inpatient records between January 1, 2014, and June 30, 2019, from the electronic medical record (EMR) system of Rui Jin Hospital, Luwan Branch, Shanghai, China. We used data from before July 1, 2018, as the development set for internal validation and the remaining data as the test set for external validation. We included patient demographics, preoperative lab results, and free-text preoperative notes as our features. We used word-embedding techniques to encode text information, and we trained the LASSO (least absolute shrinkage and selection operator) model, random forest model, gradient boosting decision tree (GBDT) model, convolutional neural network (CNN) model, and self-attention network model using the combined data. Surgeons manually scored the NNIS risk index values.

Results: For internal bootstrapping validation, CNN yielded the highest mean area under the receiver operating characteristic curve (AUROC) of 0.889 (95% CI 0.886-0.892), and the paired-sample *t* test revealed statistically significant advantages as compared with other models (*P*<.001). The self-attention network yielded the second-highest mean AUROC of 0.882 (95% CI 0.878-0.886), but the AUROC was only numerically higher than the AUROC of the third-best model, GBDT with text embeddings (mean AUROC 0.881, 95% CI 0.878-0.884, *P*=.47). The AUROCs of LASSO, random forest, and GBDT models using text embeddings were statistically higher than the AUROCs of models not using text embeddings (*P*<.001). For external validation, the self-attention network yielded the highest AUROC of 0.879. CNN was the second-best model (AUROC 0.878), and GBDT with text embeddings was the third-best model (AUROC 0.872). The NNIS risk index scored by surgeons had an AUROC of 0.651.

Conclusions: Our AMRAMS based on EMR data and deep learning methods—CNN and self-attention network—had significant advantages in terms of accuracy compared with other conventional machine learning methods and the NNIS risk index. Moreover, the semantic embeddings of preoperative notes improved the model performance further. Our models could replace the NNIS risk index to provide personalized guidance for the preoperative intervention of SSIs. Through this case, we offered an easy-to-implement solution for building multimodal RAMs for other similar scenarios.

(JMIR Med Inform 2020;8(6):e18186) doi:10.2196/18186

KEYWORDS

surgical site infection; machine learning; deep learning; natural language processing; artificial intelligence; risk assessment model; routinely collected data; electronic medical record; neural network; word embedding

Introduction

Health care–associated infection (HAI) is a global patient safety problem, with surgical site infection (SSI) being one of the most common types of HAI [1-4]. The incidences of SSI for inpatients undergoing operations are 2%-5% in the United States [5], 2%-10% in Europe [6-9], and 4%-6% in China [10-13]. SSIs increase mortality and long-term disabilities, prolong hospital length of stay (LOS), and raise health care costs [1,5,11]. In China, SSIs prolong hospital LOS by 6-23 days and increase medical costs by US $2000-$6000 per patient, with the additional cost for one SSI patient needing to be offset by the medical revenue from 13 surgical patients [11].

In 2016, the World Health Organization recommended a large perioperative care bundle of interventions for preventing SSIs, which includes perioperative oxygen inhalation; maintenance of normal body temperature; maintenance of adequate glucose and circulating volume; use of sterile drapes, surgical gowns, wound-protector devices, and antimicrobial-coated sutures; provision of incisional wound irrigation; and prophylactic negative-pressure wound therapy [14]. However, the quality of evidence for most of these recommended interventions remains low. When we do not know whether these interventions are effective enough, using several interventions together is reasonable, and may even have a summation effect, for reducing the risk of SSI as much as possible. However, the shortcomings of bundle interventions are also apparent: they will consume large amounts of medical resources, especially when we strictly implement the recommendations of the guideline. Thus, data-driven guidance for personalized intervention is key to creating more effective SSI prevention and control programs.

Many institutions have developed risk assessment models (RAMs) focusing on SSIs to help surgeons preoperatively identify high-risk patients and guide clinical interventions. The most widely used traditional RAM is the National Nosocomial Infections Surveillance (NNIS) risk index [15], which is a scoring system ranging from 0 to 3. An American Society of Anesthesiologists (ASA) preoperative assessment score higher than 2; contaminated, dirty, or infected operation; and prolonged operation duration each account for 1 point in the NNIS risk index scoring system. The risk of SSI increases from 1.5% to 13.0% as the score goes up. Obviously, the three variables are easy to calculate, but are not enough to describe the characteristics of high-risk patients. To remedy these deficiencies, Mu et al included more patient- and hospital-specific variables and developed improved RAMs for each procedure under the 39 National Healthcare Safety Network (NHSN) procedure categories using stepwise logistic regression [16]. They trained these procedure-specific models using 849,659 patient records, from 2006 to 2008, from the NHSN database. Each model used 12-15 variables, including patient demographics, anesthesia, surgery, hospital settings, and NNIS risk index factors. The overall area under the receiver operating characteristic curve (AUROC) of the model reached 0.67, higher than the AUROC of the NNIS risk index, which is 0.60. The biggest problem with their RAM is that 39 different models need to be deployed together to achieve full functionality, which is cumbersome for clinical use. In a later study, Grant et al developed another RAM, using routinely collected surveillance data from three national networks in Switzerland, France, and England [17]. They trained a logistic regression model using 46,320 colorectal surgery records from 2007 to 2017 and compared it with the previous model developed by Mu et al. In their dataset, the new model, with an AUROC of 0.65, outperformed the model developed by Mu et al. Their model was easy to use but was limited to colorectal surgery only. Meanwhile, in the absence of a high-accuracy RAM, van Walraven and Musselman developed a logistic regression model based on 362,431 clinical data points from the National Surgical Quality Improvement Program [18]. The AUROC of this model reached 0.80. However, it required the users to provide large amounts of medical history information, such as ASA score, NNIS risk index, tumor history, medication history, and operation history. These variables are not always well structured in many electronic medical record (EMR) systems, and without the support of automatic extraction, completing evaluations based on this model undoubtedly consumed large amounts of time. Therefore, a gap still exists between current preoperative RAMs and the ideal RAM, which is generalized, accurate, and easy to use or deploy.

With the widespread use of hospital information systems and EMR systems in medical institutions, we can now use massive clinical data to build RAMs. In addition to structured data, we can also use natural language processing and deep learning technology to parse semantics from unstructured clinical text data and save time for manual extraction of text information. Many researchers have developed surveillance models using data from EMRs to automatically help infection control staff efficiently identify SSIs among massive medical records and have achieved high accuracy [19-22]. However, these models used not only preoperative information but also surgical, postoperative, and antibiotic information. Thus, they cannot be used to guide preoperative intervention.

To fill in the gap, we aimed to provide a solution in the form of the Artificial intelligence–based Multimodal Risk Assessment Model for Surgical site infection (AMRAMS) for inpatients undergoing operations using routinely collected data from the EMR system of a general hospital in China. We believed that structured data, such as patient demographics and preoperative lab results, and free-text data, such as preoperative notes that record diagnoses and scheduled surgical information, would both help to identify high-risk patients. Thus, we planned to combine various machine learning, deep learning, and semantic representation technologies; validate the discriminations of multimodal implementations internally and externally; and compare them with the NNIS risk index score. We tested the following hypotheses: (1) AMRAMS, with various implementations, would more accurately identify high-risk patients than the old-fashioned NNIS risk index is capable of doing, (2) semantic information from preoperative notes would improve the model performance, and (3) deep learning implementations would outperform conventional machine learning implementations.

Methods

Source of Data

The Rui Jin Hospital, Luwan Branch, affiliated with the Shanghai Jiao Tong University School of Medicine, is a nonprofit academic medical center based in Huangpu District, Shanghai, China. The hospital has a total of 526 beds, of which 89 are in general surgery, 33 are in gynecology, 27 are in orthopedics, and 38 are in urology. The surgical staff performs more than 4000 operations annually. About 300 of these cases are emergency patients. We retrieved inpatient records that each had only one operation record during the hospital stay and a discharge record between January 1, 2014, and June 30, 2019, from the EMR system of Rui Jin Hospital, Luwan Branch. We used data from before July 1, 2018, as the development set for model training, hyperparameter tuning, and internal validation; we used the remaining data as the test set for external validation. The data usage of patient records for this study had been reviewed and approved by the ethics committee of the Rui Jin Hospital, Luwan Branch.

Participants and Features

We included adult patients only and excluded patients under the age of 18 years, patients with missing operation information (ie, timestamp of operation, whether or not theirs was an emergency operation, and type of anesthesia), and patients with missing demographic information (ie, gender and age).

We used both structured and unstructured preoperative clinical data from the EMR as our modeling features for this study. *Preoperative* was defined as the last record before the timestamp of the operation start time. Structured data included the following:

1. Patient demographics: age (years), gender (male or female), body height (cm), body weight (kg), and type of insurance (insured or noninsured).
2. Routine blood examination: white blood cell count (number × 10^9/L), proportion of neutrophils (%), proportion of lymphocytes (%), proportion of monocytes (%), proportion of eosinophils (%), proportion of basophils (%), lymphocyte count (number × 10^9/L), monocyte count (number × 10^9/L), eosinophil count (number × 10^9/L), red blood cell count (number × 10^{12}/L), hemoglobin concentration (g/L), mean corpuscular volume (fL), mean corpuscular hemoglobin (g/L), mean corpuscular hemoglobin concentration (g/L), and platelet count (number × 10^9/L).
3. Coagulation function examination: prothrombin time (sec), international normalized ratio, fibrinogen concentration (g/L), activated partial thromboplastin time (sec), thrombin time (sec), and d-dimer concentration (mg/L).
4. Liver and kidney function examination: total bilirubin concentration (μmol/L), direct bilirubin concentration (μmol/L), indirect bilirubin concentration (μmol/L), total bile acid concentration (μmol/L), alanine transaminase concentration (IU/L), aspartate aminotransferase concentration (IU/L), total protein concentration (g/L), albumin concentration (g/L), urea nitrogen concentration (mmol/L), creatinine concentration (μmol/L), uric acid concentration (μmol/L), and blood glucose concentration (mmol/L).
5. Plasmic electrolyte examination: potassium concentration (mmol/L), sodium concentration (mmol/L), calcium concentration (mmol/L), phosphorus concentration (mmol/L), and magnesium concentration (mmol/L).
6. Structured data elements from admission notes: current smoking status (true or false) and marital history (married, unmarried, or divorced).
7. Structured data elements from preoperative notes: emergency operation (true or false) and type of anesthesia (general anesthesia, total intravenous anesthesia, spinal anesthesia, epidural anesthesia, nerve block, or local anesthesia).
8. Preoperative LOS: the number of inpatient days between admission and operation.

Unstructured data included the free-text portion of the preoperative notes. Preoperative notes usually contain descriptions about preoperative diagnosis, operation name, indication, complications, and preventive measures. Table MA1-1 in Multimedia Appendix 1 shows two examples of preoperative notes from the development set.

Outcome

According to the No. 48 Decree issued by the Ministry of Health of the People's Republic of China in 2006 [23], the infection prevention and control department of the hospital should be responsible for the regular surveillance, analysis, and feedback on the epidemic situation of SSIs and their related risk factors. SSIs includes superficial incisional infection, deep incisional infection, and organ-space infection. In the Rui Jin Hospital, Luwan Branch, the staff of the infection prevention and control department manually identify SSIs via patient chart reviews and collect mandatory data for hospital administrators and government reporting after patient discharge. In this study, all the included patient records were reviewed by the infection prevention and control department. We categorized patient records with SSI identifications reported by the infection prevention and control department as positive samples. Likewise, we categorized patient records without SSI identifications as negative samples.

Data Preprocessing

In this study, we used routinely collected clinical data from the EMR system. Thus, outliers and missing data were common and inevitable. For outlier adjustment, we first discarded patient records with invalid age values (eg, age >120 years old). To detect the outliers, we implemented a two-stage algorithm based on the random-effects model adjusted for age and sex proposed by Welch et al [24] for all continuous features. We used an absolute standardized residual of more than 5 as a cutoff for outlier detection and manually reviewed all the outliers suggested by the model. If the outliers violated medical knowledge, we tried to correct the values via a chart review. If no information could be gained from the chart review, we considered the outliers as missing values. For missing data, we generated missing-value indicators [25] (ie, binary dummy features indicating whether the values of the original features were missing) and conducted mean imputation for continuous features and mode imputation for binary or categorical features. To make the results of model validation truly reflect real-world performance, we performed outlier detection and adjustment only on the development set. For the convenience of model training and optimization, we performed one-hot encoding for all the binary or categorical features and feature normalization for all features.

Model

Overview

Using the fastText algorithm, we first generated word embeddings based on a large Chinese corpus for further text-information encoding [26]. The fastText algorithm is an unsupervised neural network algorithm that learns distributional embeddings of semantic representation based on subword information for each word from the corpus. We then proposed both conventional machine learning methods and deep learning methods to predict the risk of SSI based on preoperative EMR data. Because we expected the distribution of the labels to be extremely imbalanced, we passed a positive sample weight (ie, the ratio of the number of non-SSIs to the number of SSIs) to the loss function (ie, cross-entropy) during the model training. We used the mini-batch gradient descent and backpropagation technique to update the parameters of the networks and set the AdamW algorithm as optimizer [27]. Furthermore, we used a random search method based on five-fold cross-validation and early stopping, if necessary, to find the optimized hyperparameters for each model.

Word Embedding

Our Chinese corpus contained approximately 4.1 GB of data from the Chinese Wikipedia, downloaded from the linguatools website [28], and approximately 96.9 MB of data from A-hospital, a Chinese medical Wikipedia website [29]. After we removed punctuations and numbers from the corpus, we used Jieba, version 0.41 [30], a Chinese text-segmentation tool, with a medical dictionary to segment the corpus into word sequences. Using the skipgram model of the fastText algorithm [31], we then trained 128-dimensional word embeddings using the preprocessed word sequences. We set the minimum size of the subwords as two and the maximum size as five. We left the other parameters at their default values.

Conventional Machine Learning Method

The conventional machine learning methods analyzed in this study included LASSO (least absolute shrinkage and selection operator) logistic regression with L1 penalty, random forest, and gradient boosting decision tree (GBDT), implemented by the XGBoost framework [32]. Because these models can be trained only by using tabular data, we first encoded the texts of the preoperative notes into the text embeddings. We segmented each text into a word sequence using Jieba and transformed it into a sequence of word embeddings, which is represented as follows:

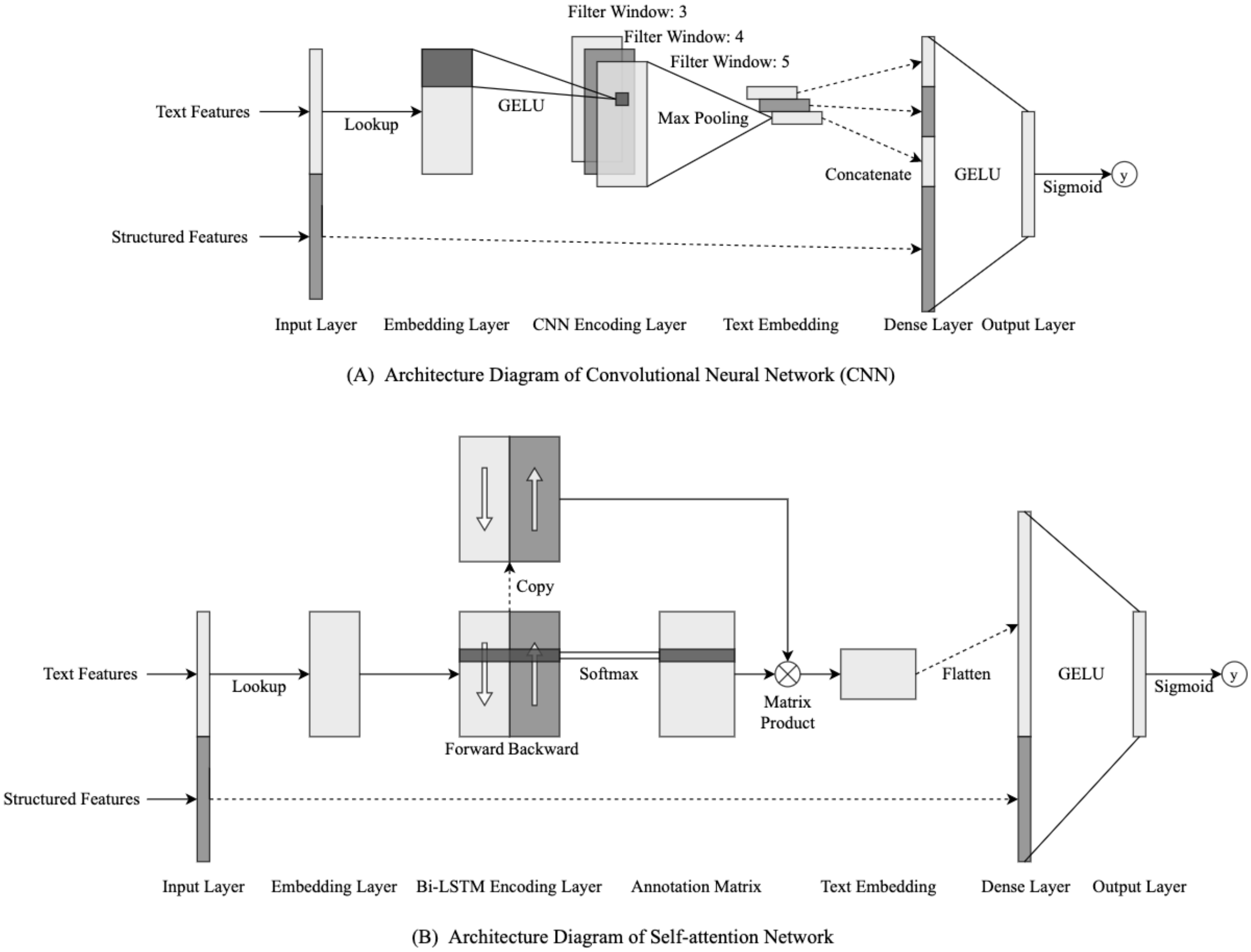
$$T = [t_1,t_2,t_3,...,t_n] \quad \mathbf{(1)}$$

Here, t_i is a 128-dimensional word-embedding vector of the i-th word in a sequence of length n, and T is an n-by-128 embedding matrix. We then pooled the embedding matrix into a 128-dimensional vector using the max-pooling method, which took the maximum value among the n words for each feature of the word embeddings. We concatenated the pooled vectors with the structured feature vectors and fed them into the four models for training.

Deep Learning Methods

The deep learning methods analyzed in this study included a convolutional neural network (CNN) and a self-attention network. In this study, we used CNN and self-attention structures to encode text information on an end-to-end basis. Figure 1 shows the architecture of both models. Before being fed into the models, text data were transformed into n-by-128 embedding matrix T. Here, n was the padding length decided by the upper boundary of the 1.5-IQR rule based on the distribution of word sequence lengths in the development set. If the actual length of the sequence was less than n, we added zero-padding to the left side of the sequence. If the actual length was more than n, we tailored the left side of the sequence to suit the padding length.

Figure 1. Deep learning network architecture diagrams. Bi-LSTM: bidirectional long short-term memory; GELU: Gaussian Error Linear Unit.



(A) Architecture Diagram of Convolutional Neural Network (CNN)

(B) Architecture Diagram of Self-attention Network

For the CNN, we referred to the structure proposed by Kim [33] and applied convolutional kernel operation on the embedding matrix T, which is represented as follows:

$$c^j_i = GELU(w^j_c \cdot t_{i:I+h-1} + b) \quad \textbf{(2)}$$

Here, c^j_i is the output value of the j-th convolutional channel of filter window i, $t_{i:i+h-1}$ is the word-embedding sequence from the i-th word to the $(i + h - 1)$-th word, h is the size of the filter window, w^j_c is the weight vector for each word embedding in the filter window of the j-th channel, b is the bias item, and *GELU* (Gaussian Error Linear Unit) [34] is the active function; the original paper used *ReLU* (Rectified Linear Unit). The filter window slides from the first word to the last one. Let m represent the number of channels for the convolutional kernel, and let n represent the length of the text; then we have an $(n - h + 1)$-by-m convolutional feature matrix:

$$C = [c_1, c_2, c_3, \ldots, c_{n-h+1}] \quad \textbf{(3)}$$

We used filter windows of three, four, and five; generated three convolutional feature matrices; applied max-pooling operation on these matrices; and concatenated the three pooled vectors and the structured feature vector together. The entire vector was then passed into a fully connected dense layer using *GELU* as an active function and, finally, a *sigmoid* output layer.

For the self-attention network, we referred to the structure proposed by Lin et al [35]. The embedding matrix T was first passed into a bidirectional long short-term memory (Bi-LSTM) layer. Let m represent the number of the hidden units for both forward and backward long short-term memory (LSTM), and let n represent the length of the word sequence. We obtained an n-by-$2m$ hidden state feature matrix:

$$H = [h_1, h_2, h_3, \ldots, h_n] \quad \textbf{(4)}$$

We then generated an attention matrix based on the hidden state feature matrix. According to the original paper, this process was similar to passing the hidden state feature matrix into two bias-free, fully connected layers using *tanh* as the first active function and *softmax* as the second function:

$$A = softmax(W_1 \cdot tanh(W_2 \cdot H^T)) \quad \textbf{(5)}$$

Here, W_2 is a d_1-by-$2m$ weight matrix, where d_1 is the hidden unit number of the first layer and W_2 is a d_2-by-d_1 weight matrix, where d_2 is the hidden unit number of the second layer. We obtained a d_1-by-$2m$ text-embedding matrix by using the following:

$$M = A \cdot H \quad \textbf{(6)}$$

We flattened the embedding matrix into a $d_1 \times 2m$-dimensional vector, concatenated it with the structured feature vector, and passed the entire vector into the dense layer.

Evaluation

We evaluated the discrimination capacities of the conventional models, with or without text embeddings, and of the deep learning models. We assessed internal validity using a bootstrapping procedure of 100 iterations based on the development set. In each iteration, we trained the model using the sample data (ie, sampling with replacement) and tested the AUROC on the out-of-bag data; the data were not sampled in the iteration. We calculated the average AUROC and 95% CI for each model and performed paired-sample *t* tests to compare the performance among the models.

To obtain a realistic estimation of the model performance, we assessed external validity on the holdout test set. We trained the models using the entire development set, tested the full training performances on the test set, and compared them with the performance of the NNIS risk index scored by surgeons before an operation. Furthermore, we calculated the sensitivity and specificity based on the test result and decided on the cutoff point using Youden’s index method [36].

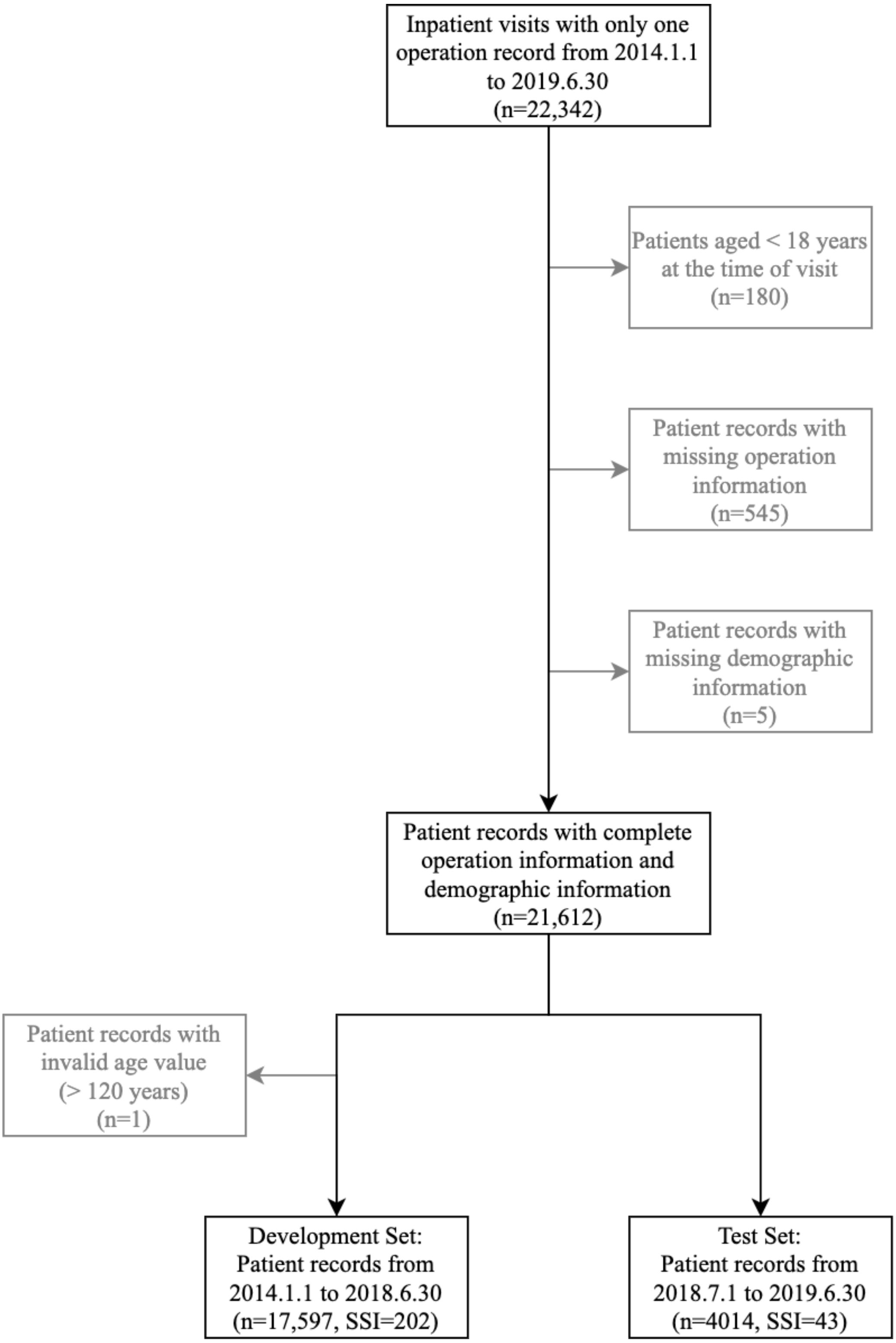
We used scikit-learn, version 0.22.1, via Python, version 3.7.4 (Python Software Foundation), to build the LASSO model and random forest model; we used XGBoost, version 0.9, via Python to build the GBDT model; and we used PyTorch, version 1.4.0, via Python to build the CNN and self-attention network. We performed all the statistical analyses using R, version 3.6.1 (The R Foundation), and considered a two-sided *P* value of <.05 as statistically significant.

Results

Patient Characteristics

We included a total of 21,611 inpatient records from January 1, 2014, to June 30, 2019. Of these records, 13,293 (61.51%) were from female patients and 8318 (38.49%) were from male patients with a median age of 54.3 years (IQR 44-65); 8375 (38.75%) were from the department of general surgery; 5903 (27.31%) were from the department of urology; 4649 (21.51%) were from the department of gynecology; and 2684 (12.42%) were from the department of orthopedics. According to the distributions of the International Classification of Diseases, Tenth Revision (ICD-10) code of operation and diagnosis that were retrieved after patient discharge, the patients received surgical treatment mainly for genitourinary system diseases, neoplasms, and digestive system diseases; the main types of operations were urinary system surgery, digestive system surgery, female reproductive system surgery, and endocrine system surgery. Overall, the incidence of SSIs in our dataset was 1.13% (244/21,611). The assigned sample size of the development set was 17,597 and that of the test set was 4014. The missing-data rates of the included variables ranged from 0% to 70.9% in the development set and from 0% to 72.8% in the test set. The variables with missing-data rates of more than 20% came from liver and kidney function examination, plasmic electrolyte examination, and d-dimer measurement. A slight difference was observed in the missing-data rate of each variable between the development set and the test set. Among them, the variables with the largest differences in the missing-data rates came from the electrolyte examinations (ie, calcium, phosphorous, and magnesium), with the rate differences reaching 8.0%. Figure 2 shows the selection process for the patient records and Table MA1-2 in Multimedia Appendix 1 shows the patient characteristics. We released a portion of the raw data in Multimedia Appendix 2, and the data dictionary of the raw data is located in the Data Description section of Multimedia Appendix 1.

Figure 2. Flowchart of the selection process for patient records. Gray boxes show records that were excluded due to patients not meeting inclusion criteria and records containing outliers or missing data. SSI: surgical site infection.


Inpatient visits with only one operation record from 2014.1.1 to 2019.6.30 (n=22,342)
Patients aged < 18 years at the time of visit (n=180)
Patient records with missing operation information (n=545)
Patient records with missing demographic information (n=5)
Patient records with complete operation information and demographic information (n=21,612)
Patient records with invalid age value (> 120 years) (n=1)
Development Set: Patient records from 2014.1.1 to 2018.6.30 (n=17,597, SSI=202)
Test Set: Patient records from 2018.7.1 to 2019.6.30 (n=4014, SSI=43)

Hyperparameters and Training

We selected the optimal hyperparameters for each model based on the results of five-fold cross-validation. For LASSO, we used an L1 penalty of 0.01 when using text embeddings and 0.003 when not using text embeddings. For random forest with text embeddings, we used 300 trees, a maximum depth of 18, and maximum features of 0.6. For random forest without text embeddings, we used 1000 trees, a maximum tree depth of 4, and maximum features of 0.6. For GBDT with text embeddings, we used a learning rate (η) of 0.01, a maximum tree depth of 24, a subsample of 0.6, a column sample of 0.65, a gamma of 0.3, and 61 iterations. For GBDT without text embeddings, we used a learning rate (η) of 0.003, a maximum tree depth of 4, a subsample of 0.65, a column sample of 0.8, a gamma of 0, and 132 iterations. For the CNN, we used a learning rate (η) of

0.0001; an L2 penalty of 3; a word-embedding layer dropout rate of 0; CNN filter windows of three, four, and five with 256 feature maps (ie, channels) each; a dropout rate of 0.35; a fully connected layer with 128 feature maps; a dropout rate of 0.5; and 18 epochs. For the self-attention network, we used a learning rate (η) of 0.0001, an L2 penalty of 0.03, a word-embedding layer dropout rate of 0.5, a Bi-LSTM with 256 feature maps (ie, hidden nodes) each, a dropout rate of 0.45, an attention network with 256 feature maps (ie, hidden nodes) on the first layer and 64 for the second layer, a fully connected layer with 128 feature maps, a dropout rate of 0, and 19 epochs. We set the padding length for deep learning to 244. Hyperparameters not mentioned in this section were left at their default values.

Model Performances

Table 1 lists the performances of the models in terms of both internal and external validation, and Figure 3 shows the receiver operating characteristic (ROC) curves of the top five models based on full training and NNIS risk index. For internal validation, CNN yielded the highest mean AUROC of 0.889 (95% CI 0.886-0.892), and the paired-sample *t* test (see Multimedia Appendix 1, Table MA1-3) revealed statistically significant advantages (*P*<.001) compared with the other models. The self-attention network yielded the second-highest mean AUROC of 0.882 (95% CI 0.878-0.886). However, the AUROC of the self-attention network was only numerically higher than the AUROC of the third-best model—GBDT with text embeddings (mean AUROC 0.881, 95% CI 0.878-0.884)—and did not exhibit statistical significance (*P*=.47). The AUROCs of the machine learning models using text embeddings were statistically higher than the AUROCs of the models not using text embeddings (*P*<.001). For external validation, the self-attention network yielded the highest AUROC of 0.879. CNN was the second-best model (AUROC 0.878), and GBDT with text embeddings was the third-best model (AUROC 0.872). The NNIS risk index scored by surgeons had an AUROC of 0.651, which was remarkably lower than that of any other model in our study. Based on the external validation, we could still observe a trend with the text embeddings improving the model performances in the external validation. All the models had lower AUROC scores in internal validation than in external validation (ie, mean AUROC).

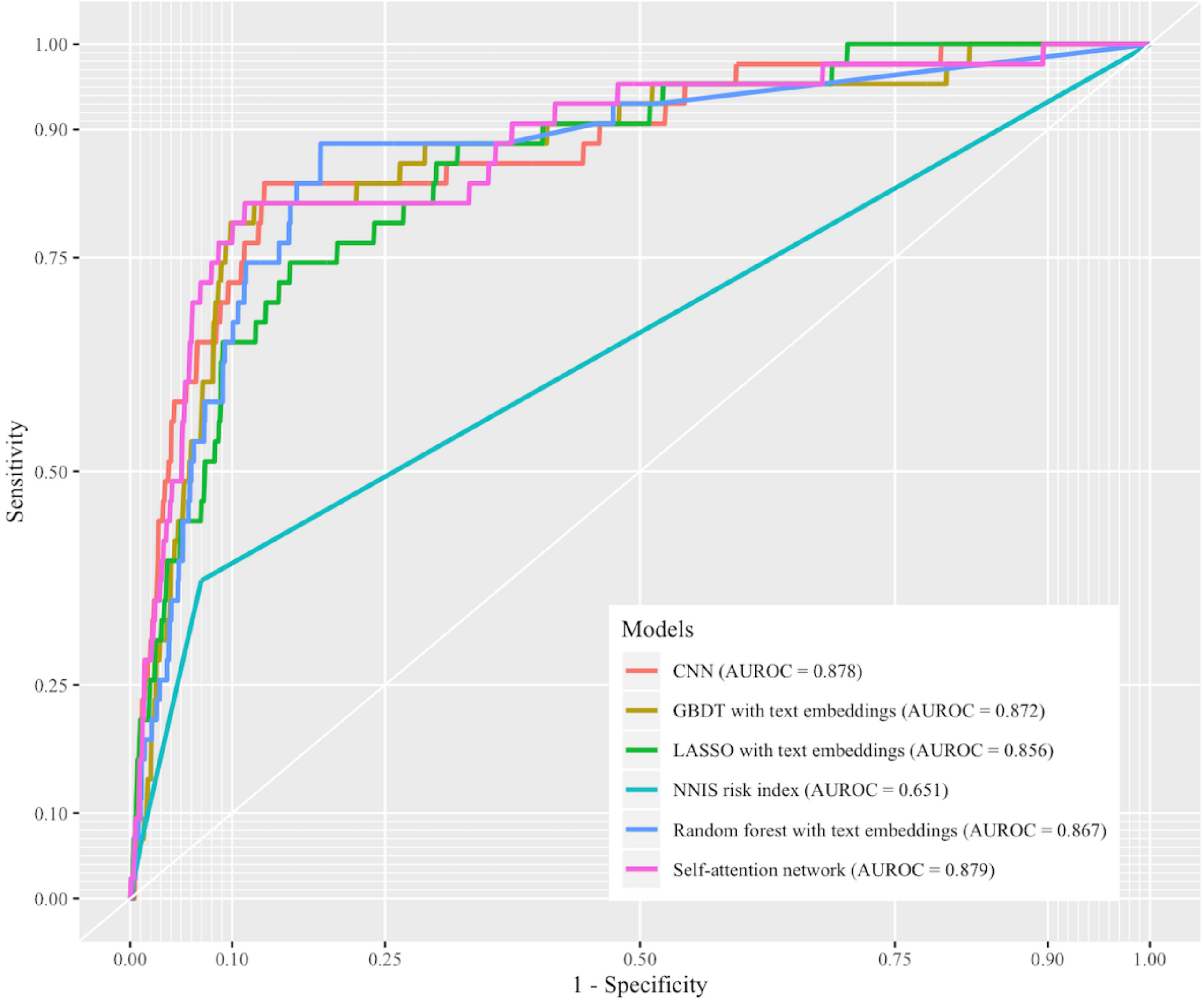
Table 1. Model performances.

Model and text embedding	Area under the receiver operating characteristic curve (AUROC)		Sensitivity[a] (full training)	Specificity[a] (full training)
	Bootstrapping, mean (95% CI)	Full training		
Least absolute shrinkage and selection operator (LASSO)				
With text embedding	0.870 (0.867-0.874)	0.856	0.744	0.844
Without text embedding	0.856 (0.852-0.860)	0.816	0.674	0.842
Random forest				
With text embedding	0.877 (0.873-0.880)	0.867	0.884	0.813
Without text embedding	0.846 (0.842-0.850)	0.772	0.558	0.871
Gradient boosting decision tree (GBDT)				
With text embedding	0.881 (0.878-0.884)	0.872	0.791	0.902
Without text embedding	0.838 (0.834-0.843)	0.782	0.605	0.858
Convolutional neural network (CNN)	0.889 (0.886-0.892)	0.878	0.837	0.869
Self-attention	0.882 (0.878-0.886)	0.879	0.814	0.888
National Nosocomial Infections Surveillance (NNIS) risk index	N/A[b]	0.651	0.372	0.930

[a]The optimal cutoff point was identified using Youden’s index method.

[b]N/A: not applicable.

Figure 3. The receiver operating characteristic (ROC) curves of the top five models based on full training and National Nosocomial Infections Surveillance (NNIS) risk index. AUROC: area under the receiver operating characteristic curve; CNN: convolutional neural network; GBDT: gradient boosting decision tree; LASSO: least absolute shrinkage and selection operator.



Feature Analysis

Both deep learning models—CNN and self-attention network—performed better than other models in our validations. However, the deep learning models were black boxes and hard to explain. To further explore the correlations between the selected features and the occurrence of SSIs, we conducted a population-level feature analysis for the structured features and a case-level analysis for the text embeddings. The population-level analysis explored the correlations by comparing the normalized coefficient for each feature from LASSO without text embeddings; the coefficients were based on the data after normalization. For case-level analysis, we referenced the idea from local interpretable model-agnostic explanations [37]. For each case, we fixed the structured features and generated new word sequences by randomly removing words from the raw sequence and dummy binary vectors that indicated whether the word in a certain position was removed or not. We generated 10,000 new sequences for each case, combined them with the structured features, passed them to the deep learning models, and obtained prediction scores. We then fitted a LASSO regression model—with an L1 penalty of 0.01—that uses dummy binary vectors as features and prediction scores as targets. The coefficients of the LASSO regression model indicated the relative contributions of the words to the prediction scores in a case.

Figure 4 shows the features with nonzero coefficients and their coefficients for the population-level analysis. We could observe that preoperative LOS, marital history, anesthesia type, gender, age, results of routine blood examination, coagulation function examination, and many missing-value indicators had remarkable impacts on the model. Among them, patients with prolonged preoperative LOS, patients with missing AST results, married patients, older patients, and patients with missing body weight information had higher risks of SSI. Patients with higher hemoglobin, female patients, patients with missing magnesium results, patients that had received total intravenous anesthesia, and patients with missing marital histories had lower risks of SSI.

Figure 4. The normalized coefficients of the features in the LASSO (least absolute shrinkage and selection operator) model without text embeddings. ALB: albumin; APTT: activated partial thromboplastin time; AST: aspartate aminotransferase; CA: calcium; DBIL: direct bilirubin; EA: epidural anesthesia; GLU: blood glucose; HGB: hemoglobin; INR: international normalized ratio; K: potassium; LOS: length of stay; LYMPH: lymphocyte; MCH: mean corpuscular hemoglobin; MCV: mean corpuscular volume; MG: magnesium; MONO: monocyte; NA: sodium; PP: phosphorus; SA: spinal anesthesia; TBIL: total bilirubin; TIVA: total intravenous anesthesia; TP: total protein; TT: thrombin time; UA: uric acid.

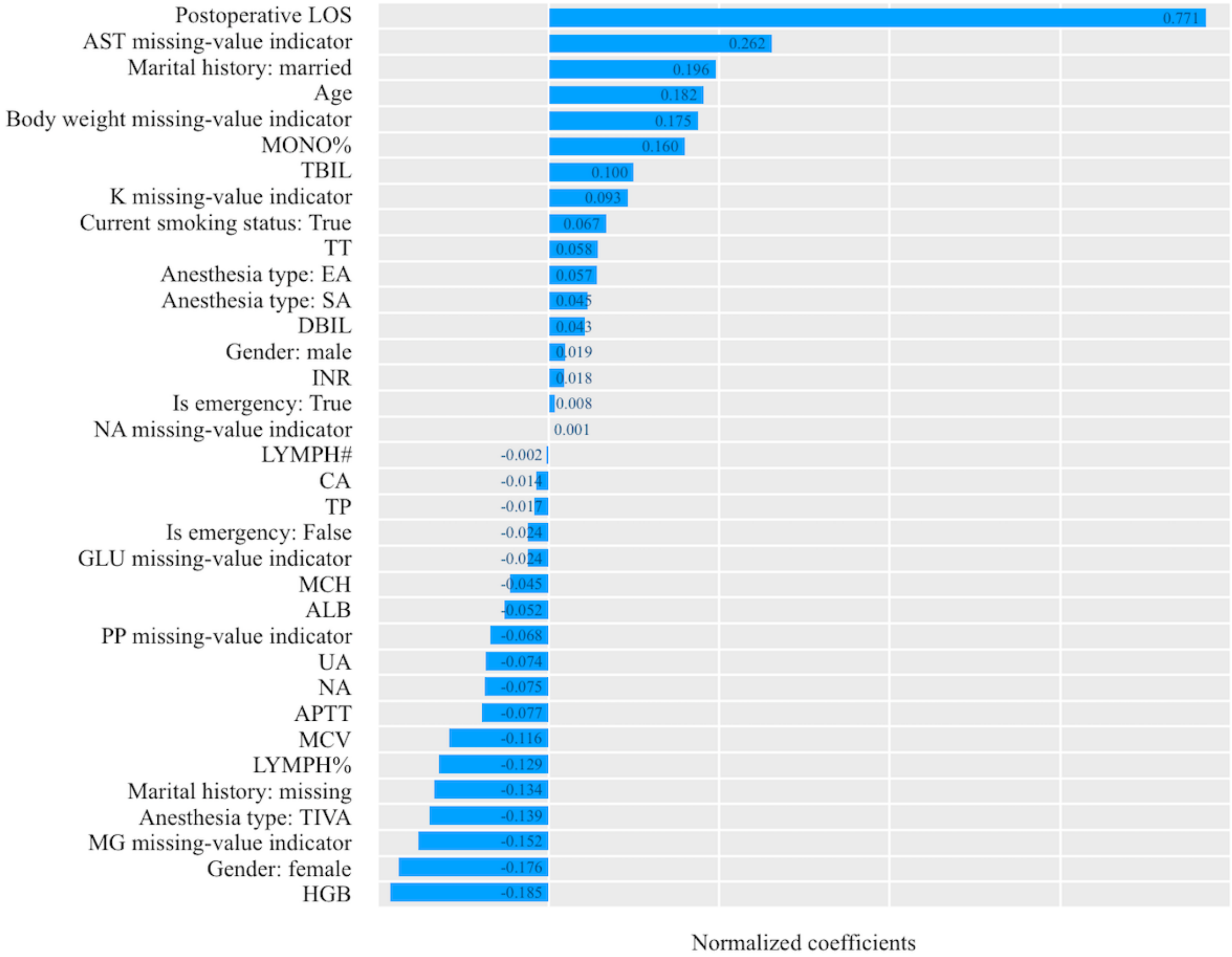
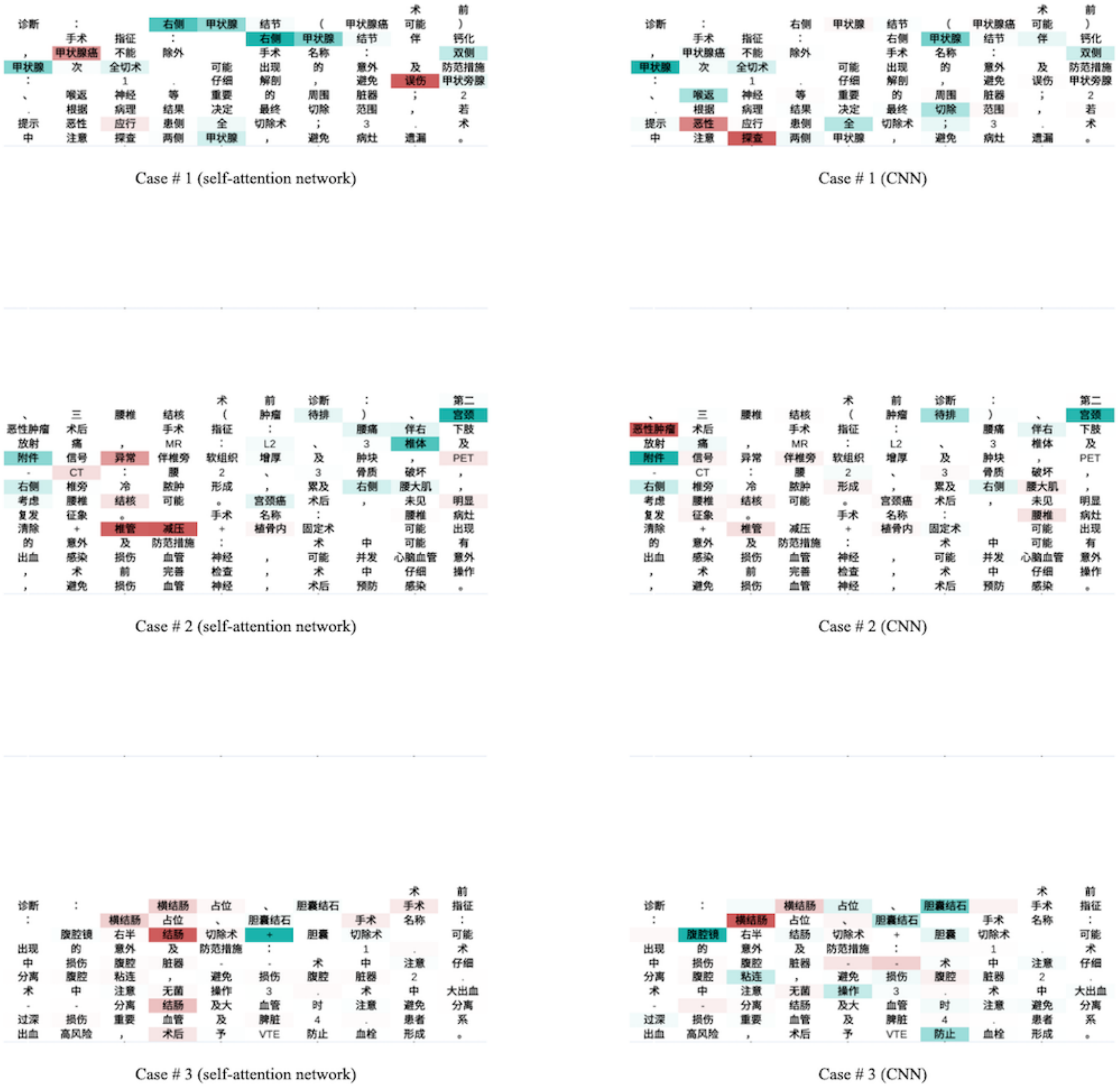


Figure 5 shows the heatmaps of the word contributions to the CNN and attention prediction scores for three preoperative note cases (see Table MA1-1 in Multimedia Appendix 1 for the full-text translation), with green being a negative coefficient (ie, protective factor) and red being a positive coefficient (ie, risk factor). The deeper the color, the higher the absolute value. Among the three cases, we could observe that terms like “甲状腺 (thyroid),” “宫颈 (uterine neck),” “附件 (accessory),” “椎体 (centrum),” “腹腔镜 (laparoscope),” and “胆囊结石 (gallstone)” were associated with lower risk of SSI, and terms like “甲状腺癌 (thyroid cancer),” “恶性 (malignant),” “结核 (tuberculosis),” “恶性肿瘤 (malignant tumor),” “结肠 (colon),” and “横结肠 (transverse colon)” were associated with higher risk of SSI.

Figure 5. The heatmaps of the word contributions on three preoperative note cases. CNN: convolutional neural network.



Case # 1 (self-attention network)

Case # 1 (CNN)

Case # 2 (self-attention network)

Case # 2 (CNN)

Case # 3 (self-attention network)

Case # 3 (CNN)

Discussion

Principal Findings

In this study, we found that SSI RAMs based on clinical data from an EMR system and modern machine learning techniques could identify high-risk patients more accurately than the old-fashioned NNIS risk index is capable of doing. Notably, the vectorial embedding of preoperative notes, whether generated using a simple max-pooling method or using a deep learning method, improved the performance of the model further without any handcrafted feature engineering. The multimodal deep learning models that produced end-to-end feature representations automatically through convolutional kernels, LSTM, or attention mechanisms outperformed the traditional machine learning models, such as LASSO, random forest, or GBDT. Thus, our AMRAMS using a CNN or a self-attention network could replace the NNIS risk index in providing personalized guidance for the preoperative intervention of SSI. At the same time, our study provided an easy-to-implement solution to building a multimodal RAM for similar scenarios based on both structured and Chinese text data. Because we used routinely collected preoperative data only, such as the results of routine blood tests and clinical notes, additional manual data collection and clinician evaluation was no longer necessary for achieving high accuracy.

Many factors could explain the advantages of the deep learning AMRAMS. First, we used more objective quantitative features of the patients. The NNIS risk index contained only three elements, of which the ASA score was a subjective feature provided by anesthesiologists. The model developed by Mu’s team, on the other hand, utilized 12-15 features and included the ASA score [16]. The model developed by Grant’s team utilized seven features and also included the ASA score [17]. The model developed by van Walraven and Musselman utilized

53 features and included not only subjective features, such as the ASA score, the NNIS risk index score, and dyspnea evaluation, but also 33 manually extracted variables from the medical history of the patient [18]. Meanwhile, our model included 47 objective features, among which age, gender, body weight, anesthesia type, emergency operation, preoperative hemoglobin, glucose, and LOS have proven to be related to the occurrences of SSI [12,38,39]. All these features were the results of routine preoperative blood tests and were automatically extracted from the EMR system using SQL (Structured Query Language) query script.

Second, we used sufficient information about the operative procedures and risk prevention from the preoperative notes via the fastText embeddings and the network structures. Many studies on English text classification have demonstrated that machine learning using fastText embeddings had better performance than those of the algorithms using bag-of-words, n-grams, TF-IDF (term frequency–inverse document frequency), or word2vec embeddings [26,40]. The semantics of many words, especially in the Chinese language, depend on the subwords or characters they contain. The fastText algorithm generated the semantic embeddings based on the internal structures of words, which best suit the characteristics of natural language. Moreover, our deep learning models enabled end-to-end learning: both fastText embeddings and hidden nodes of the network could be further fine-tuned simultaneously according to the specified targets during the learning process. To encode text-level semantics, we tried both convolutional kernel and attention mechanisms. These network structures could automatically represent n-grams and long-term dependency information, which helped the deep learning models gain better performance than those of conventional machine learning models (ie, logistic regression, naïve-Bayes, and support-vector machine) trained using the top of max-pooling embeddings [33,35]. Because of the complexity of the deep learning models, we were not able to precisely identify the decision mechanisms of the text semantics. However, according to the heatmaps of case-level word contribution, the potentially essential keywords helped identify SSIs in the form of distributed representation without any other handcrafted feature engineering or manual feature extraction. Potentially essential keywords included those suggesting endoscopic surgery, such as “laparoscope,” and “gallstone”; those suggesting clean surgery, such as “thyroid” and “centrum”; those suggesting colon surgery, such as “colon” and “transverse colon”; and those suggesting complex operation and prolonged operation time, such as “malignant,” “tuberculosis,” and “malignant tumor.”

Third, we tried many algorithmic techniques to avoid overfitting. For example, we used the L2 penalty, dropout, and early-stopping techniques. These techniques ensured the generalization ability of the model on different patient data to a certain extent.

Although we verified the effectiveness of the deep learning AMRAMS through both internal and external verification, many limitations still exist. The first limitation came from the training labels. The follow-up period of SSI by the infection prevention and control department was limited only to the hospitalization, which meant some of the SSIs that occurred after discharge might have been ignored. Although surgeons would conduct a careful examination of the incision before patient discharge, the occurrence of SSIs outside secondary care could not be completely eliminated. This bias would cause our model to underestimate the risk of patients developing SSIs.

The second limitation came from the patient population. Our dataset came from one medical site, and the time span of data collection was about 5 years, in which changes in patient population distribution, surgical procedures, and SSI prevention education and measures were inevitable. In our study, the internal validation results of the models were not completely consistent with the external verification results, which implied this point. If our model was to be applied to clinical practice, regular validation and update would be necessary.

The third limitation came from the missing values. We observed that many variables in our dataset had high missing rates, and many missing-value indicators contributed greatly to the model. In general, the missing data in the EMR system were not missing at random and were caused mainly by two reasons: inability to perform the measurement or a lack of indication to perform the measurement. For example, missing body weight information might indicate that the patient was unable to stand upright (eg, paralyzed) in order to measure the weight, whereas missing blood tests and liver function tests might suggest that the patient was healthy and young. In our study, we were not able to evaluate the potential influence of the missing data, because speculating the reason behind each missing value was complex and trivial. From a perspective of research, we could try to model the probability of missingness using other observed variables and conduct sensitivity analyses, which stimulate various missing patterns based on the predicted probability distributions, to evaluate the influences of missing data in our future studies. However, the ideal solutions to missing-data problems are still improving data quality and integrity in EMRs or developing less-biased imputation methods based on the patterns of the missing values in the patient records, the conditions of the patients, and the behaviors of the physicians.

The fourth limitation came from the models. We did not observe the attention mechanisms on the top of Bi-LSTM providing a great benefit over the convolutional kernels as claimed by Lin et al in their paper [35], probably because of the limited sizes of the training samples relative to the model parameters. Thus, we considered both self-attention and CNN as the best solutions in our current task. Because of the limitations of computing resources (ie, GPU [graphics processing unit] instances), we did not apply other state-of-the-art language models, such as BERT (bidirectional encoder representations from transformers) and its derivatives [41,42], to encode text information.

The fifth limitation came from the feature analysis. In this study, we only explored the correlations between the selected features and the occurrences of SSIs via the LASSO models, with detailed epidemic investigations and causal inferences remaining beyond the scope of this study. Moreover, because the LASSO models were trained using incomplete data and were not adjusted for potential confounders, the statistical inferences would be biased and the results would be hard to explain.

Future studies will focus on four points. First, we will try various new language models with deep transformers; encode various text information types, such as admission records, progress notes, and surgical records; and evaluate the models' performance. Second, we will confirm the effectiveness of our AMRAMS among multiple medical sites. Third, we will embed the AMRAMS into the EMR system and evaluate whether it can ultimately help reduce the occurrence of SSIs and optimize medical decision making. Fourth, our multimodal RAM solution could be validated for many other similar scenarios, such as syndromic or notifiable disease surveillance, adverse event monitoring, or ICD-10 coding support, in which both structured and free-text features would contribute to the judgement of final outcomes.

Conclusions

Our artificial intelligence–based multimodal risk assessment models for SSI based on EMR data and deep learning methods had significant advantages in terms of accuracy, compared with other conventional machine learning methods and the NNIS risk index. The semantic embeddings of clinical notes, whether generated using a simple max-pooling method or a deep learning method, improved the model performance further without any handcrafted feature engineering. Our models could replace the NNIS risk index to provide personalized guidance for the preoperative intervention of SSI. Through this case, we offered an easy-to-implement solution for building multimodal RAMs for similar scenarios, based on both structured and free-text data. Future studies should validate the generalization, reproducibility, and clinical impact in other medical settings.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The data description of the raw data and additional tables.
[DOCX File , 31 KB - medinform_v8i6e18186_app1.docx]

Multimedia Appendix 2
A portion of the patient records raw data.
[XLSX File (Microsoft Excel File), 4627 KB - medinform_v8i6e18186_app2.xlsx]

References

1. Cassini A, Plachouras D, Eckmanns T, Abu Sin M, Blank H, Ducomble T, et al. Burden of six healthcare-associated infections on European population health: Estimating incidence-based disability-adjusted life years through a population prevalence-based modelling study. PLoS Med 2016 Oct;13(10):e1002150 [FREE Full text] [doi: 10.1371/journal.pmed.1002150] [Medline: 27755545]
2. Lake JG, Weiner LM, Milstone AM, Saiman L, Magill SS, See I. Pathogen distribution and antimicrobial resistance among pediatric healthcare-associated infections reported to the National Healthcare Safety Network, 2011-2014. Infect Control Hosp Epidemiol 2018 Jan;39(1):1-11 [FREE Full text] [doi: 10.1017/ice.2017.236] [Medline: 29249216]
3. Hansen S, Schwab F, Zingg W, Gastmeier P, The Prohibit Study Group. Process and outcome indicators for infection control and prevention in European acute care hospitals in 2011 to 2012 - Results of the PROHIBIT study. Euro Surveill 2018 May;23(21):1-10 [FREE Full text] [doi: 10.2807/1560-7917.ES.2018.23.21.1700513] [Medline: 29845929]
4. Leaper D, Ousey K. Evidence update on prevention of surgical site infection. Curr Opin Infect Dis 2015 May;28(2):158-163. [doi: 10.1097/QCO.0000000000000144] [Medline: 25692267]
5. Waltz PK, Zuckerbraun BS. Surgical site infections and associated operative characteristics. Surg Infect (Larchmt) 2017;18(4):447-450. [doi: 10.1089/sur.2017.062] [Medline: 28448197]
6. Cossin S, Malavaud S, Jarno P, Giard M, L'Hériteau F, Simon L, ISO-RAISIN Steering Committee. Surgical site infection after valvular or coronary artery bypass surgery: 2008-2011 French SSI national ISO-RAISIN surveillance. J Hosp Infect 2015 Dec;91(3):225-230. [doi: 10.1016/j.jhin.2015.07.001] [Medline: 26321674]
7. Pollard TC, Newman JE, Barlow NJ, Price JD, Willett KM. Deep wound infection after proximal femoral fracture: Consequences and costs. J Hosp Infect 2006 Jul;63(2):133-139. [doi: 10.1016/j.jhin.2006.01.015] [Medline: 16621145]
8. Leaper D, Nazir J, Roberts C, Searle R. Economic and clinical contributions of an antimicrobial barrier dressing: A strategy for the reduction of surgical site infections. J Med Econ 2010;13(3):447-452. [doi: 10.3111/13696998.2010.502077] [Medline: 20653399]
9. Gheorghe A, Moran G, Duffy H, Roberts T, Pinkney T, Calvert M. Health utility values associated with surgical site infection: A systematic review. Value Health 2015 Dec;18(8):1126-1137 [FREE Full text] [doi: 10.1016/j.jval.2015.08.004] [Medline: 26686800]
10. Wang Z, Chen J, Wang P, Jie Z, Jin W, Wang G, et al. Surgical site infection after gastrointestinal surgery in China: A multicenter prospective study. J Surg Res 2019 Aug;240:206-218. [doi: 10.1016/j.jss.2019.03.017] [Medline: 30986636]

11. Zhou J, Ma X. Cost-benefit analysis of craniocerebral surgical site infection control in tertiary hospitals in China. J Infect Dev Ctries 2015 Mar 19;9(2):182-189 [FREE Full text] [doi: 10.3855/jidc.4482] [Medline: 25699493]
12. Fan Y, Wei Z, Wang W, Tan L, Jiang H, Tian L, et al. The incidence and distribution of surgical site infection in mainland China: A meta-analysis of 84 prospective observational studies. Sci Rep 2014 Oct 30;4:6783 [FREE Full text] [doi: 10.1038/srep06783] [Medline: 25356832]
13. Xiao Y, Shi G, Zhang J, Cao J, Liu L, Chen T, et al. Surgical site infection after laparoscopic and open appendectomy: A multicenter large consecutive cohort study. Surg Endosc 2015 Jul;29(6):1384-1393. [doi: 10.1007/s00464-014-3809-y] [Medline: 25303904]
14. Allegranzi B, Zayed B, Bischoff P, Kubilay NZ, de Jonge S, de Vries F, WHO Guidelines Development Group. New WHO recommendations on intraoperative and postoperative measures for surgical site infection prevention: An evidence-based global perspective. Lancet Infect Dis 2016 Dec;16(12):e288-e303. [doi: 10.1016/S1473-3099(16)30402-9] [Medline: 27816414]
15. Culver DH, Horan TC, Gaynes RP, Martone WJ, Jarvis WR, Emori TG, et al. Surgical wound infection rates by wound class, operative procedure, and patient risk index. National Nosocomial Infections Surveillance System. Am J Med 1991 Oct 16;91(3B):152S-157S. [doi: 10.1016/0002-9343(91)90361-z] [Medline: 1656747]
16. Mu Y, Edwards JR, Horan TC, Berrios-Torres SI, Fridkin SK. Improving risk-adjusted measures of surgical site infection for the National Healthcare Safety Network. Infect Control Hosp Epidemiol 2011 Oct;32(10):970-986. [doi: 10.1086/662016] [Medline: 21931247]
17. Grant R, Aupee M, Buchs NC, Cooper K, Eisenring M, Lamagni T, et al. Performance of surgical site infection risk prediction models in colorectal surgery: External validity assessment from three European national surveillance networks. Infect Control Hosp Epidemiol 2019 Sep;40(9):983-990. [doi: 10.1017/ice.2019.163] [Medline: 31218977]
18. van Walraven C, Musselman R. The Surgical Site Infection Risk Score (SSIRS): A model to predict the risk of surgical site infections. PLoS One 2013;8(6):e67167 [FREE Full text] [doi: 10.1371/journal.pone.0067167] [Medline: 23826224]
19. Bucher BT, Ferraro JP, Finlayson SR, Chapman WW, Gundlapalli AV. Use of computerized provider order entry events for postoperative complication surveillance. JAMA Surg 2019 Apr 01;154(4):311-318 [FREE Full text] [doi: 10.1001/jamasurg.2018.4874] [Medline: 30586132]
20. Pindyck T, Gupta K, Strymish J, Itani KM, Carter ME, Suo Y, et al. Validation of an electronic tool for flagging surgical site infections based on clinical practice patterns for triaging surveillance: Operational successes and barriers. Am J Infect Control 2018 Feb;46(2):186-190. [doi: 10.1016/j.ajic.2017.08.026] [Medline: 29031434]
21. Grundmeier RW, Xiao R, Ross RK, Ramos MJ, Karavite DJ, Michel JJ, et al. Identifying surgical site infections in electronic health data using predictive models. J Am Med Inform Assoc 2018 Sep 01;25(9):1160-1166. [doi: 10.1093/jamia/ocy075] [Medline: 29982511]
22. Colborn KL, Bronsert M, Amioka E, Hammermeister K, Henderson WG, Meguid R. Identification of surgical site infections using electronic health record data. Am J Infect Control 2018 Nov;46(11):1230-1235. [doi: 10.1016/j.ajic.2018.05.011] [Medline: 29907448]
23. The Ministry of Health of the People's Republic of China, No. 48 Decree. URL: http://www.gov.cn/ziliao/flfg/2006-07/25/content_344886.htm [accessed 2020-05-19]
24. Welch C, Petersen I, Walters K, Morris RW, Nazareth I, Kalaitzaki E, et al. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. Pharmacoepidemiol Drug Saf 2012 Jul;21(7):725-732. [doi: 10.1002/pds.2270] [Medline: 22052713]
25. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995 Dec 15;142(12):1255-1264. [doi: 10.1093/oxfordjournals.aje.a117592] [Medline: 7503045]
26. Santos I, Nedjah N, de Macedo Mourelle L. Sentiment analysis using convolutional neural network with fastText embeddings. In: Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). 2017 Nov 08 Presented at: 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI); November 8-10, 2017; Arequipa, Peru p. 1-5. [doi: 10.1109/la-cci.2017.8285683]
27. Loshchilov I, Hutter F. arXiv. 2017 Nov. Decoupled weight decay regularization URL: https://ui.adsabs.harvard.edu/abs/2017arXiv171105101L [accessed 2020-05-24]
28. linguatools. Wikipedia monolingual corpora URL: https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/ [accessed 2020-05-19]
29. A-hospital. URL: http://www.a-hospital.com/ [accessed 2020-05-19]
30. GitHub. Jieba: Chinese text segmentation URL: https://github.com/fxsjy/jieba [accessed 2020-05-19]
31. Bojanowski P, Grave E, Joulin A, Mikolov T. arXiv. 2016 Jul. Enriching word vectors with subword information URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160704606B [accessed 2020-05-24]
32. Chen T, He T. Higgs boson discovery with boosted trees. In: Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS): 2014 Workshop in High Energy Physics and Machine Learning (HEPML). 2015 Presented at: 28th Conference on Neural Information Processing Systems (NIPS): 2014 Workshop in High Energy Physics and Machine Learning (HEPML); December 8-13, 2014; Montreal, Canada p. 69-80 URL: http://www.jmlr.org/proceedings/papers/v42/chen14.pdf

33. Kim Y. arXiv. 2014 Aug. Convolutional neural networks for sentence classification URL: https://ui.adsabs.harvard.edu/abs/2014arXiv1408.5882K [accessed 2020-05-24]
34. Hendrycks D, Gimpel K. arXiv. 2016 Jun. Gaussian Error Linear Units (GELUs) URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160608415H [accessed 2020-05-24]
35. Lin Z, Feng M, Nogueira dos Santos C, Yu M, Xiang B, Zhou B. arXiv. 2017 Mar. A structured self-attentive sentence embedding URL: https://ui.adsabs.harvard.edu/abs/2017arXiv170303130L [accessed 2020-05-24]
36. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. Epidemiology 2005 Jan;16(1):73-81. [doi: 10.1097/01.ede.0000147512.81966.ba] [Medline: 15613948]
37. Tulio Ribeiro M, Singh S, Guestrin C. arXiv. 2016 Feb. Why should I trust you? Explaining the predictions of any classifier URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160204938T [accessed 2020-05-24]
38. Gong S, Guo H, Zhou H, Chen L, Yu Y. Morbidity and risk factors for surgical site infection following cesarean section in Guangdong Province, China. J Obstet Gynaecol Res 2012 Mar;38(3):509-515. [doi: 10.1111/j.1447-0756.2011.01746.x] [Medline: 22353388]
39. Gomila A, Carratalà J, Biondo S, Badia JM, Fraccalvieri D, Shaw E, VINCat Colon Surgery Group. Predictive factors for early- and late-onset surgical site infections in patients undergoing elective colorectal surgery. A multicentre, prospective, cohort study. J Hosp Infect 2018 May;99(1):24-30. [doi: 10.1016/j.jhin.2017.12.017] [Medline: 29288776]
40. Joulin A, Grave E, Bojanowski P, Mikolov T. arXiv. 2016 Jul. Bag of tricks for efficient text classification URL: https://ui.adsabs.harvard.edu/abs/2016arXiv160701759J [accessed 2020-05-24]
41. Devlin J, Chang MW, Lee K, Toutanova K. arXiv. 2018 Oct. BERT: Pre-training of deep bidirectional transformers for language understanding URL: https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D [accessed 2020-05-24]
42. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. arXiv. 2019 Sep. ALBERT: A lite BERT for self-supervised learning of language representations URL: https://ui.adsabs.harvard.edu/abs/2019arXiv190911942L [accessed 2020-05-24]

Abbreviations

AMRAMS: Artificial intelligence–based Multimodal Risk Assessment Model for Surgical site infection
ASA: American Society of Anesthesiologists
AUROC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
Bi-LSTM: bidirectional long short-term memory
CNN: convolutional neural network
EMR: electronic medical record
GBDT: gradient boosting decision tree
GELU: Gaussian Error Linear Unit
GPU: graphics processing unit
HAI: health care–associated infection
ICD-10: International Classification of Diseases, Tenth Revision
LASSO: least absolute shrinkage and selection operator
LOS: length of stay
LSTM: long short-term memory
NHSN: National Healthcare Safety Network
NNIS: National Nosocomial Infections Surveillance
RAM: risk assessment model
ReLU: Rectified Linear Unit
ROC: receiver operating characteristic
SQL: Structured Query Language
SSI: surgical site infection
TF-IDF: term frequency–inverse document frequency

Edited by G Eysenbach; submitted 11.02.20; peer-reviewed by H Zhang, A Ćirković; comments to author 28.03.20; revised version received 15.04.20; accepted 19.04.20; published 15.06.20.

Please cite as:
Chen W, Lu Z, You L, Zhou L, Xu J, Chen K
Artificial Intelligence–Based Multimodal Risk Assessment Model for Surgical Site Infection (AMRAMS): Development and Validation Study
JMIR Med Inform 2020;8(6):e18186
URL: http://medinform.jmir.org/2020/6/e18186/
doi:10.2196/18186
PMID:32538798

Original Paper

Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation

Tianzhou Yang[1*], MD; Li Zhang[1*], PhD; Liwei Yi[2], MD; Huawei Feng[1], PhD; Shimeng Li[1], PhD; Haoyu Chen[2], MD; Junfeng Zhu[1], PhD; Jian Zhao[1], MD; Yingyue Zeng[1], PhD; Hongsheng Liu[1,3,4], PhD

[1]School of Life Science, Liaoning University, Shenyang, China

[2]School of Information, Liaoning University, Shenyang, China

[3]Research Center for Computer Simulating and Information Processing of Bio-macromolecules of Shenyang, Liaoning University, Shenyang, China

[4]Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning, Shenyang, China

[*]these authors contributed equally

Corresponding Author:
Hongsheng Liu, PhD
School of Life Science
Liaoning University
No. 66, Chongshan Middle road
Shenyang, 110036
China
Phone: 86 024 62202280
Fax: 86 024 62202280
Email: liuhongsheng@lnu.edu.cn

Abstract

Background: Early diabetes screening can effectively reduce the burden of disease. However, natural population–based screening projects require a large number of resources. With the emergence and development of machine learning, researchers have started to pursue more flexible and efficient methods to screen or predict type 2 diabetes.

Objective: The aim of this study was to build prediction models based on the ensemble learning method for diabetes screening to further improve the health status of the population in a noninvasive and inexpensive manner.

Methods: The dataset for building and evaluating the diabetes prediction model was extracted from the National Health and Nutrition Examination Survey from 2011-2016. After data cleaning and feature selection, the dataset was split into a training set (80%, 2011-2014), test set (20%, 2011-2014) and validation set (2015-2016). Three simple machine learning methods (linear discriminant analysis, support vector machine, and random forest) and easy ensemble methods were used to build diabetes prediction models. The performance of the models was evaluated through 5-fold cross-validation and external validation. The Delong test (2-sided) was used to test the performance differences between the models.

Results: We selected 8057 observations and 12 attributes from the database. In the 5-fold cross-validation, the three simple methods yielded highly predictive performance models with areas under the curve (AUCs) over 0.800, wherein the ensemble methods significantly outperformed the simple methods. When we evaluated the models in the test set and validation set, the same trends were observed. The ensemble model of linear discriminant analysis yielded the best performance, with an AUC of 0.849, an accuracy of 0.730, a sensitivity of 0.819, and a specificity of 0.709 in the validation set.

Conclusions: This study indicates that efficient screening using machine learning methods with noninvasive tests can be applied to a large population and achieve the objective of secondary prevention.

(JMIR Med Inform 2020;8(6):e15431) doi:10.2196/15431

KEYWORDS

type 2 diabetes; screening; non-invasive attributes; machine learning

Introduction

Diabetes is a heterogeneous metabolic disorder that is characterized by the presence of hyperglycemia due to impairment of insulin secretion, defective insulin action, or both [1]. The high blood glucose level caused by diabetes not only affects the heart, eyes, kidneys, and nerves but also is associated with increased rates of cancer, physical and cognitive disabilities [2-4], tuberculosis [5,6], and depression [7]; these conditions are associated with high health care costs [8,9]. For patients with type 2 diabetes, the risks of death and cardiovascular events are 2-4 times greater than in the general population [10]. Due to the aging population, lifestyle changes, and interrelated rapid unplanned urbanization, the prevalence of diabetes is quickly increasing worldwide [11]. According to the latest International Diabetes Federation Diabetes Atlas, there were approximately 420 million people aged 20-79 years with diabetes worldwide in 2017, and this number is expected to rise to 629 million in 2045. Furthermore, approximately 50% of diabetes patients are undiagnosed [12]. Patients with type 2 diabetes who are within target ranges for 5 risk factor variables, namely glycated hemoglobin levels, systolic and diastolic blood pressure, albuminuria, smoking, and low-density lipoprotein cholesterol levels, appear to have little or no excess risk of death, myocardial infarction, or stroke compared with the general population [13]. Therefore, developing an appropriate method to screen people without clinical symptoms is necessary and practical; such a screening method could reduce health care costs and patient mortality and improve patients' quality of life through earlier clinic-based management.

Generally, traditional screening projects are based on studies in epidemiology, such as the ADDITION trial study [14] and the Ely study [15]. These screening studies cost hundreds of thousands of dollars and require the collaboration of many people. With the emergence and development of machine learning, researchers have started to pursue more flexible and efficient methods to screen or predict type 2 diabetes. Han et al [16] trained a type 2 diabetes diagnosis model with features mainly consisting of blood tests such as hemoglobin $A1_c$ and total cholesterol, yielding a precision of 0.942 and a recall of 0.939. Maniruzzaman et al [17,18] trained a type 2 diabetes prediction model using Pima Indian data with plasma glucose features; they obtained an accuracy of 81.97% and an area under the curve (AUC) of 0.93. A machine learning–based framework was also developed to identify patients with type 2 diabetes in the clinic with electronic health records, showing an AUC of 0.98 with more than 110 clinical features [19]. Zou et al [20] used principal component analysis and minimum redundancy maximum relevance to reduce the dimensionality and achieve the best accuracy in their model (0.81) in addition to using fasting blood sugar as the main feature. Many of the abovementioned studies achieved high prediction performance with blood tests; however, none of them used only noninvasive attributes to predict type 2 diabetes. Chung et al [21] developed a model to screen prediabetes using support vector machines with only noninvasive features, such as age, sex, and family history of diabetes, and they obtained an AUC of 0.76 in the external test data; however, further exploration and optimization are needed to improve type 2 diabetes screening models that only use noninvasive features.

To better screen potential patients with type 2 diabetes, further delay disease progression, control relative complications, and improve human health, in this paper, type 2 diabetes screening machine learning models and conforming easy ensemble models were built that require only an individual noninvasive test, combined with data from body measurements and questionnaires, to predict type 2 diabetes based on the National Health and Nutrition Examination Survey (NHANES) database, thus avoiding blood tests and clinic visits. Inexpensive screening of people who have type 2 diabetes without obvious symptoms may lead to secondary prevention.

Methods

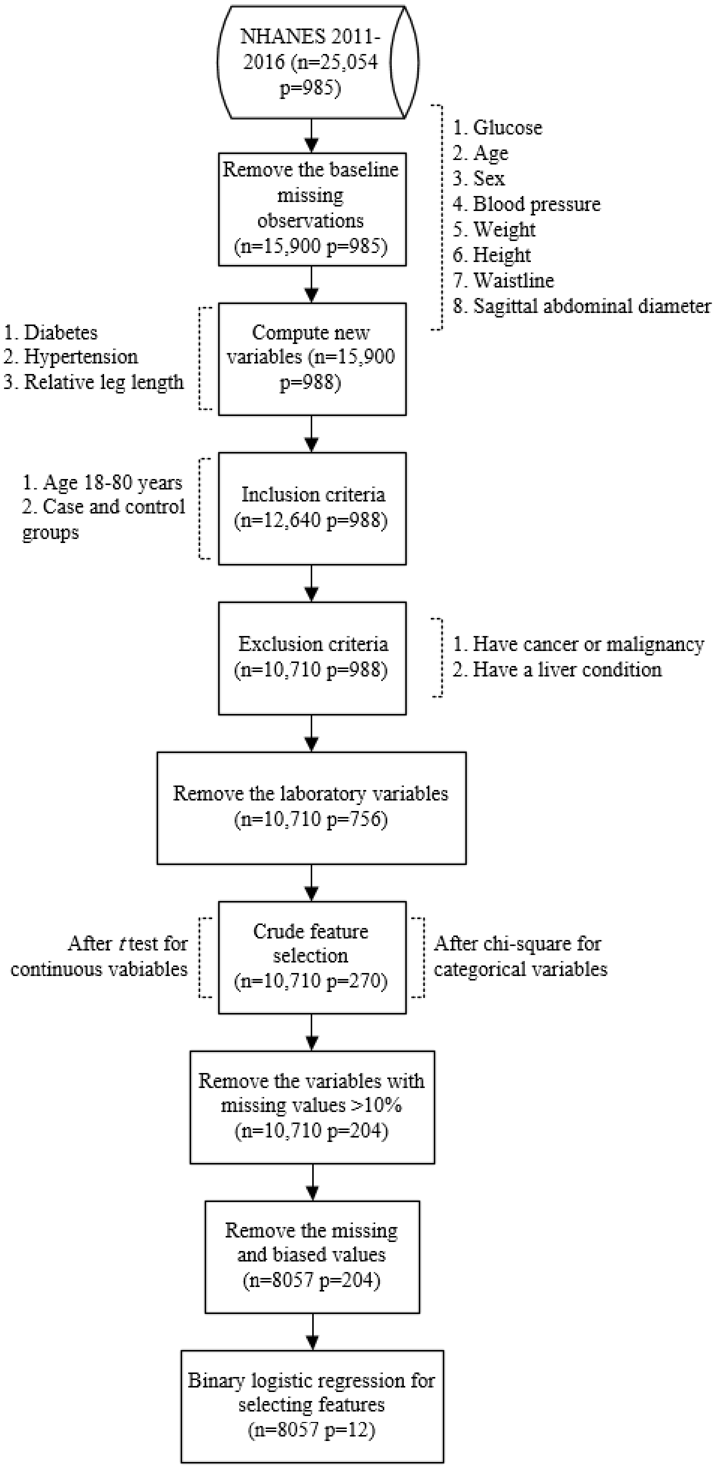
Analysis

The data were analyzed with R version 3.3.1 for Linux with the R packages dplyr, caret (Classification And REgression Training) [22], randomForest [23], pROC [24], e1071 [25], gplots, unbalanced [26], epiDisplay, and MASS. The Delong test for 2 correlated receiver operating characteristic (ROC) curves was used to determine the effects of the easy ensemble methods; a *P* value <0.05 was considered significant (2-sided). The work protocol consisted of 5 steps: data cleaning, sample selection, chosen features, model training, and validation.

Data

The data were obtained from the NHANES database. The detailed steps of data cleaning and feature selection are shown in Figure 1. First, before all the NHANES data were processed, the database contained 25,054 samples from 2011 to 2016 with 985 features. Second, data samples with missing observations for baseline variables, such as blood glucose, age, sex, height, and weight, were removed. Third, 3 new variables were computed, namely diabetes (whether a person has diabetes: 1=yes, 0=no), hypertension (whether a person has hypertension: 1=yes, 0=no) and relative leg length. The case group was defined as having fasting blood glucose levels ≥7.0 millimoles per liter, and the fasting blood glucose levels in the control group were <6.1 mmol/L [1]. Hypertension was defined according to the American Heart Association criteria as systolic blood pressure ≥130 millimeters of mercury or diastolic blood pressure ≥80 mm Hg obtained on more than 2 occasions [27]. The relative leg length was the ratio of the upper leg length to the height multiplied by 100 [28]. Fourth and fifth, we set the inclusion and exclusion criteria to control for bias. The inclusion criteria were as follows: patients aged 18-80 years from the case and control groups. The following exclusion criteria were employed: patients with cancer, due to the positive association between hyperglycemia and cancer [29], and patients with liver conditions, because liver conditions can also influence blood glucose levels [30]. These individuals were excluded because they are traditionally asymptomatic and their blood glucose levels are not representative of the study population. After the data processing steps (1-5), 10,710 observations and 988 features without type 2 diabetes were left for analysis.

Figure 1. The data cleaning and feature selection process. Note that the feature selection process was run only in the NHANES 2011-2014 dataset. n: number of cases. p: number of features.


NHANES 2011-2016 (n=25,054 p=985)
Remove the baseline missing observations (n=15,900 p=985)
1. Glucose
2. Age
3. Sex
4. Blood pressure
5. Weight
6. Height
7. Waistline
8. Sagittal abdominal diameter
1. Diabetes
2. Hypertension
3. Relative leg length
Compute new variables (n=15,900 p=988)
1. Age 18-80 years
2. Case and control groups
Inclusion criteria (n=12,640 p=988)
Exclusion criteria (n=10,710 p=988)
1. Have cancer or malignancy
2. Have a liver condition
Remove the laboratory variables (n=10,710 p=756)
After t test for continuous vabiables
Crude feature selection (n=10,710 p=270)
After chi-square for categorical variables
Remove the variables with missing values >10% (n=10,710 p=204)
Remove the missing and biased values (n=8057 p=204)
Binary logistic regression for selecting features (n=8057 p=12)

Feature Selection

The selection of features is one of the most critical steps in model building. Thus, additional feature selection steps were taken. First, because only noninvasive features were used, the laboratory variables were deleted, and 756 features were left. Secondly, we used the *t* test to select continuous variables and the chi-square test to select the categorical variables for crude feature selection with *P*<.05; this resulted in 270 remaining features. Third, the variables whose missing values were greater than 10% were removed, leaving 204 features. Fourth, the missing and biased values (including answers in the questionnaire such as “refused” and “don’t know”) were deleted, leaving 8057 samples. Finally, forward conditional logistic regression was employed to further filter the features that were selected in the former steps with *P*<.05 only in the NHANES

2011-2014 dataset. After the feature selection process, 12 features remained. We separated the final dataset into three parts: the training set (80%, 2011-2014) with 3582 negative and 664 positive observations, the test set (20%, 2011-2014) with 895 negative and 165 positive observations, and the external validation set (2015-2016) with 2244 negative and 507 positive observations; the whole 2011-2014 data set was randomly divided into the training set and test set using the createDataPartition function in the caret package [22].

Machine Learning and the Easy Ensemble Method

In this study, binary logistic regression was used to select the risk factors for diabetes, and the linear discriminant analysis, random forest, and support vector machine methods as well as their ensemble methods were developed to classify the case and control groups according to the selected features. The linear discriminant analysis structure was based on the lda function of the R package MASS, the support vector machine structure was based on the svm function of the R package e1071, and the random forest structure was based on the rf function of the R package randomForest. The parameter adjustments of the support vector machine and random forest were applied with the R package caret. We used 80% of the 2011-2014 NHANES data for model training under 100 repeated 5-fold cross-validations. The remaining 20% of the 2011-2014 NHANES data were used as the test set, and the 2015-2016 NHANES data were reserved as the validation set for performance measurement.

Logistic Regression

As an extension of linear regression, logistic regression is a commonly used method to obtain the risk or protection factors for disease in epidemiology [31,32]. According to the experimental design, this logic function was divided into unconditional and conditional logistic regressions; according to the type of dependent variables, it was divided into binary logistic regression and multiple logistic regression. The logistic function is an effective method for classification problems and gives the odds ratio (OR) of the significance variable according to the dependent variable.

In this study, binary unconditional logistic regression was used to select the risk factors for or relative features of diabetes. In the logistic regression, the 204 attributes chosen from the *t* test and chi-square test were considered as the independent variables, and whether a person has diabetes was the dependent variable. Twelve features were left.

Linear Discriminant Analysis

Linear discriminant analysis was first introduced by Fisher [33] in 1936 to address taxonomic problems. Generally, it is a combination of analysis of variance and regression analysis. Linear discriminant analysis is based on the theory of transformation from high dimensions to low dimensions. As a classification algorithm, its theoretical basis is that the protection points of each type of data are as close to each other as possible, while the distance between different kinds of data are as far apart as possible. In this case, the classification was based on whether a person has diabetes. Therefore, the linear discriminant analysis reduced the 12 features to the 1(k–1, k=2) dimension to discriminate patients with diabetes.

Random Forest

Random forest, which is based on decision trees [34], is a well-known ensemble learning method that uses the bagging method [35]. The basic theory of the bagging method is as follows: assuming a dataset contains N observations, for example, 100 subsets can be extracted wherein every subset comprises n (n=N) observations that were sampled randomly with replacement from the original dataset, and 100 base classifiers can be built with these 100 subsets to vote for the classification of every sample in the dataset. The decision trees are the base classifier in the bagging method in the random forest. This basic algorithm can be considered as a single tree model with if-then structures. Each decision tree of the RF yields its own classification outcome and "vote," and the average of all the results is the final taxonomy.

The caret package in R was applied to search for the best parameter in the random forest with 5-fold cross-validation repeated 100 times. The number of trees was 500, and the best number of variables randomly sampled as candidates at each split was 4 after the parameter selection.

Support Vector Machine

Support vector machines [36] are among the most popular supervised learning techniques in the machine learning field. A support vector machine reflects the data to a higher-dimensional space with a kernel function. The classification mission relies on the training data, which are called support vectors. For general 2-class problems, the observations are determined by a hyperplane with the maximizing margin through the nearest support vectors.

In this study, the radial basis kernel was chosen. The caret package of R was also used to match the parameter with the best AUC performance in the support vector machine model with 5-fold cross-validation repeated 100 times. The optimal cost and gamma parameter values obtained for the model were 0.137 and 0.012, respectively.

Easy Ensemble Method

Type 2 diabetes screening is an unbalanced problem because there are fewer patients than healthy individuals. To address the unbalanced issue, we employed the easy ensemble method [37]. In short, we randomly sampled the same number of all positive observations from the negative observations and made the two groups correspond to a minor dataset in the train set. We then repeated the above step 100 times to generate 100 minor datasets. Next, we built 100 same-method models based on these datasets. Furthermore, for 5-fold cross-validation, the prevalence probability of every sample was averaged by these 100 models in every validation for both the test set and validation set.

Model Evaluation

In this article, we used the ROC curve, AUC, sensitivity, specificity, accuracy, and positive predictive value (PPV) to measure the performance of the models. The cutoff value was selected based on the maximal value of the Youden index [38] in the training set.

Results

After the data cleaning and feature selection process, the dataset included 8057 cases that were divided into three sets: 80% of the NHANES 2011-2014 data for the training set, 20% of the NHANES 2011-2014 data for the test set, and the NHANES 2015-2016 data for the validation set. After crude feature selection with the *t* test and chi-square test in the 2011-2014 NHANES dataset, logistic regression analysis was further performed to assess the related factors of type 2 diabetes; this process ensures that there will be no overfitting or generalization of the model for future patients. The 12 selected factors are shown in Table 1.

Table 1. Factors associated with diabetes used to build the models.

Feature	Crude[a] OR[b] (95% CI)	Adjusted[c] OR (95% CI)	*P* value
Age	1.05 (1.05-1.06)	1.05 (1.04-1.06)	<.001
Sex	0.82 (0.70-0.97)	0.62 (0.50-0.76)	<.001
Waistline	1.04 (1.03-1.05)	0.99 (0.97-1.01)	.27
Sagittal abdominal diameter	1.20 (1.18-1.22)	1.16 (1.09-1.24)	<.001
Relative leg length	0.70 (0.66-0.74)	0.85 (0.79-0.91)	<.001
60 second pulse	1.02 (1.01-1.02)	1.02 (1.01-1.03)	<.001
Smoking	0.74 (0.63-0.88)	1.13 (0.92-1.38)	.26
Alcohol	1.43 (1.19-1.72)	1.31 (1.04-1.66)	.02
Hypertension	3.26 (2.72-3.90)	1.02 (0.82-1.27)	.86
Family history	0.28 (0.24-0.34)	0.32 (0.26-0.39)	<.001
General health condition	2.05 (1.88-2.24)	1.59 (1.44-1.76)	<.001
Control or loss of weight	0.42 (0.35-0.51)	0.55 (0.44-0.69)	<.001

[a] Crude: 1-way logistic regression.

[b] OR: odds radio.

[c] Adjusted: multiple logistic regression.

The risk of having type 2 diabetes increases with increased age (95% CI 1.04-1.06, *P*<.001), sagittal abdominal diameter (95% CI 1.09-1.24, *P*<.001), pulse (95% CI 1.01-1.03, *P*<.001), and alcohol use (95% CI 1.04-1.66, *P*=.02) as well as poorer general health condition (95% CI 1.44-1.76, *P*<.001). In contrast, female sex, longer relative leg length, lack of type 2 diabetes family history, and control of weight are the protection factors of type 2 diabetes (95% CI 0.50-0.76, 0.79-0.91, 0.26-0.39, and 0.44-0.69, respectively; *P*<.001 in all cases). We built three different models using linear discriminant analysis, random forest, and support vector machine methods to determine type 2 diabetes risk using the training set with these noninvasive tests. Afterward, the test set and external validation set were used to measure the predictive ability of the models.

We generated six models with three different machine learning methods as well as corresponding ensemble methods in the training set. The 5-fold cross-validation results in Table 2 show that the linear discriminant analysis method yielded the best AUC compared with the random forest and support vector machine methods not only with the simple methods but also with the easy ensemble methods. However, the ensemble method improvements in the different methods are in the order of support vector machine > random forest > linear discriminant analysis. In 5-fold cross-validation, the simple linear discriminant analysis method showed 0.844 AUC, 74.1% sensitivity, 79.5% specificity, 78.7% accuracy, and 40.2% PPV; the ensemble linear discriminant analysis method showed 0.845 AUC, 79.7% sensitivity, 73.5% specificity, 74.5% accuracy, and 35.8% PPV. The simple random forest method showed 0.823 AUC, 86.2% sensitivity, 61.2% specificity, 65.1% accuracy, and 29.2% PPV; its ensemble method showed 0.834 AUC, 78.4% sensitivity, 73.2% specificity, 74.0% accuracy, and 35.2% PPV. The simple support vector machine method showed 0.808 AUC, 69.2% sensitivity, 81.1% specificity, 79.2% accuracy, and 40.5% PPV; the ensemble support vector machine method showed 0.842 AUC, 78.7% sensitivity, 74.8% specificity, 75.4% accuracy, and 36.7% PPV. The line graph in Figure 2 shows that the AUC improved with accumulation of the models, and the values remained stable after the composition of approximately 10 models.

Table 2. Average results (SD) of the 5-fold cross-validation of the models in the training set.

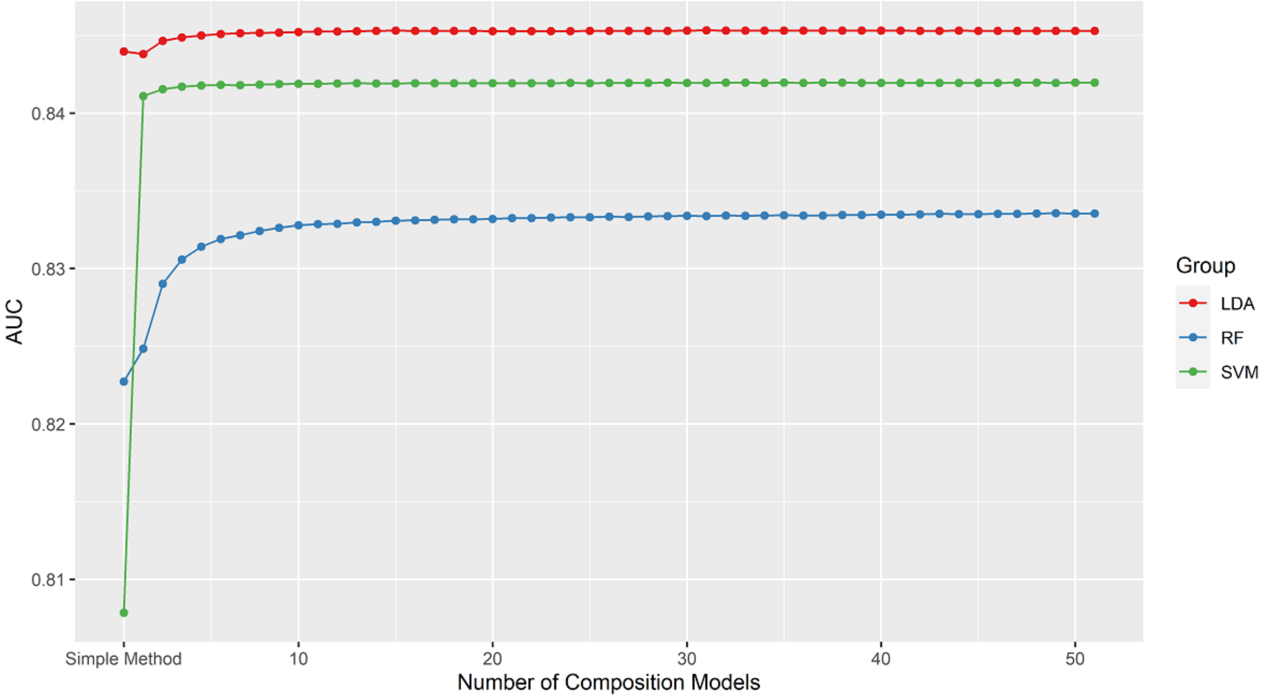
Method	AUC[a]	Sensitivity	Specificity	Accuracy	PPV[b]
Simple methods					
Linear discriminant analysis	0.844 (0.016)	0.741 (0.035)	0.795 (0.015)	0.787 (0.013)	0.402 (0.020)
Random forest	0.823 (0.016)	0.862 (0.029)	0.612 (0.019)	0.651 (0.015)	0.292 (0.011)
Support vector machine	0.808 (0.015)	0.692 (0.035)	0.811 (0.017)	0.792 (0.014)	0.405 (0.023)
Ensemble methods					
EE[c] linear discriminant analysis	0.845 (0.016)	0.797 (0.032)	0.735 (0.016)	0.745 (0.014)	0.358 (0.017)
EE random forest	0.834 (0.016)	0.784 (0.033)	0.732 (0.016)	0.740 (0.014)	0.352 (0.016)
EE support vector machine	0.842 (0.016)	0.787 (0.034)	0.748 (0.017)	0.754 (0.014)	0.367 (0.018)

[a] AUC: area under the curve.

[b] PPV: positive predictive value.

[c] EE: easy ensemble method.

Figure 2. Comparison of the top 50 models with the easy ensemble method and the simple method with different machine learning methods and 5-fold cross-validation in the training set. AUC: area under the curve. LDA: linear discriminant analysis. RF: random forest. SVM: support vector machine.



The 5-fold cross-validation indicated that the different models show reliable capability. Similarly, the AUCs of the developed models range from 0.810-0.850 in the test and validation datasets, indicating their stability and extensibility for predicting the risk of new patients with type 2 diabetes. Furthermore, when considering the performance of the easy ensemble methods in the test set (Table 3), these methods appeared to predict type 2 diabetes more efficiently than the other methods. For the random forest and support vector machine methods, the easy ensemble methods provided significantly better AUC values than the respective simple methods (absolute AUC improvement 0.014, z=3.062, *P*=.002 and 0.07, z=5.010, *P*<.001, respectively), as determined by the Delong test for two correlated ROC curves (2-sided). However, the LDA improvement was not significant (z=1.252, *P*=.21) according to the Delong test. In the validation set (Table 3), we found a similar pattern. The easy ensemble methods improved the overall predictive performance by 0.004 (z=2.734, *P*=.006) for linear discriminant analysis, 0.008 (z=2.991, *P*=.002) for random forest, and 0.037 (z=5.908, *P*<.001) for support vector machine.

The results indicate that the ensemble methods can be used to screen large populations for type 2 diabetes based on their significantly improved performance in the tests for the random forest and support vector machine methods and in the external validation set for the linear discriminant analysis, random forest, and support vector machine methods. For better and easier application of type 2 diabetes screening, a screening website based on the ensemble method has been established [39].

Table 3. Performance of the simple and ensemble methods in the text and validation sets.

Method	AUC[a]	Sensitivity	Specificity	Accuracy	PPV[b]
Test set					
Simple methods					
Linear discriminant analysis	0.864	0.697	0.829	0.808	0.429
Random forest	0.836	0.830	0.648	0.676	0.303
Support vector machine	0.796	0.630	0.864	0.827	0.460
Ensemble methods					
EE[c] linear discriminant analysis	0.867	0.758	0.777	0.774	0.385
EE random forest	0.850	0.776	0.770	0.771	0.383
EE support vector machine	0.861	0.752	0.783	0.778	0.390
Validation set					
Simple methods					
Linear discriminant analysis	0.846	0.759	0.762	0.761	0.418
Random forest	0.828	0.888	0.594	0.648	0.331
Support vector machine	0.811	0.720	0.789	0.776	0.435
Ensemble methods					
EE[c] linear discriminant analysis	0.849	0.819	0.709	0.730	0.389
EE random forest	0.836	0.813	0.713	0.731	0.390
EE support vector machine	0.848	0.824	0.714	0.734	0.394

[a] AUC: area under the curve.

[b] PPV: positive predictive value.

[c] EE: easy ensemble method.

Discussion

Comparison With Prior Work

The results of one analysis predicted that the world ranking of the number of years of life lost due to diabetes will increase from 15th to 7th [40] by 2040. The fact that type 2 diabetes damages health conditions deserves special attention. In this article, we generated type 2 diabetes screening models and applied them to a large population. Although some researchers [16-20] have studied machine learning models for screening and predicting type 2 diabetes, most of their studies focused on improving performance by selecting many features, such as blood test results, instead of considering the practical significance of cost and flexibility. In contrast, we used a noninvasive test covering demographic factors, body measurements, and questionnaire variables to build our models; this addresses the shortcomings of using invasive tests. Jai Won Chung et al [21] also adopted noninvasive features to predict prediabetes, including age, gender, family history of diabetes, hypertension, alcohol intake, BMI, smoking status, waist circumference, and physical activity; they obtained a best AUC of 0.76 in the external test data. However, the attributes they chose were relatively traditional compared with those chosen in this study; in addition, the similarities between prediabetes and healthy cases can result in lower AUC values. The validation of our models indicates that body measurements and questionnaire questions can be used to predict whether a person has type 2 diabetes. In the case of further effects resulting from high blood sugar conditions, the models can be used to screen the identified people.

Principal Results

In the feature selection process in this study, traditional analyses such as the *t* test, chi-square test, and binary logistic regression were used. We extracted unusual attributes related to type 2 diabetes, such as sagittal abdominal diameter, relative leg length, and heart rate, which were proven to be significant in similar studies [28,41,42], in addition to some common risk factors, such as age, sex, alcohol use, and family history [43,44]. Among these features, relative leg length was an interesting clue to type 2 diabetes that has not previously been used in type 2 diabetes prediction; this feature was selected by *t* test and forward conditional logistic regression. Epidemiological studies from various settings indicate that humans with shorter legs relative to their stature have higher risk for type 2 diabetes [28]. Relative leg length can be easily determined and has a strong correlation with type 2 diabetes; therefore, it may be a useful new attribute in model building or epidemiology research. With increasing adoption of this feature, our model will be more accurate and dependable.

Reliable type 2 diabetes screening models based on noninvasive tests and machine learning algorithms were established and validated in this study. All the easy ensemble methods yielded higher predictive performance (AUC≥0.85 and AUC≥0.83, respectively) in the test set and validation set than the simple methods, indicating the efficiency of the ensemble methods. Screening models based on population are always an unbalanced problem, with more negative samples and fewer positive samples in the whole dataset. In other words, the learning ability of the models is not satisfied by the positive samples. We randomly matched a negative sample for every positive sample and generated 100 base models. This type of repeated learning from the positive samples may improve the results of the models. In addition to AUC, the application of the ensemble can increase the steadiness of the performance; this was exhibited by other measurements, such as sensitivity, specificity, accuracy, and PPV. Compared with different machine learning methods, the ensemble method improvement is limited; this suggests that the dataset and features are more essential. In recent research, the results show that individuals with screen-detected type 2 diabetes were diagnosed earlier and had better outcomes than those who were clinically detected with regard to all-cause mortality, cerebrovascular disease, renal disease, and retinopathy [45]. In addition to earlier ordinal treatment, Ej et al [46] introduced a method to recover the function of islets by diet control. Regardless of treatment, quality of life improvement and decreased disease burden are important.

Limitations

There are several limitations of our research project. The World Health Organization definition of diabetes is inferior to proper diagnosis by an experienced physician; also, we cannot clearly separate type 1 diabetes from type 2 diabetes, which would cause bias because of their different epidemiological attributes. After removing the baseline missing values and executing the inclusion and exclusion criteria, there were 10,710 samples in the entire database. Additionally, 2653 missing and biased values were removed. The proportion of patients with diabetes to patients without diabetes is approximately 1:5; therefore, the increased amount of abandoned diabetes data may reduce the predictive ability of the model. Reproducibility remains doubtful given the variable demographics of the different datasets. Only a study using noninvasive features to screen for diabetes can minimize the impact of demographic changes such as those considered in large population health studies and nutrition surveys. The best PPV was only 0.435 in the validation set; this indicates that only approximately 40% of true positive samples from the people detected positively by these models were patients with type 2 diabetes. A higher false-positive value increases the financial expenses of the health care system in the beginning; however, this type of screening program can improve the overall health of the population, and earlier diagnosis can decrease the disease burden, ultimately decreasing health care expenses related to diabetes. On one hand, although the easy ensemble method [37] applied here addresses the unbalanced problem in one sense, more positive observations may yield better performance; on the other hand, the building of type 2 diabetes screening models is always an imbalanced problem when screening patients with type 2 diabetes from a large population. Therefore, we cannot solve the unbalanced problem completely. After considering all the other possible biases influencing the performance of the models, the key point is to further explore and optimize the unbalanced problem.

Conclusions

Accurate models with low-cost variables based on NHANES data for screening type 2 diabetes were established; the models performed better with the application of ensemble methods. The use of NHANES data by the models ensured a sufficient sample size, and the models can be a tool to determine the health conditions of people who were not included in the survey. Compared with prior literature, this study has certain advantages, such as noninvasive features and reliable model performance. However, we still obtained low PPV results for the unbalanced problem and could not completely solve the missing value problem. Furthermore, we can not only optimize the method by incorporating more quality data from medical schools but can also combine our study with a cohort study to achieve primary prevention.

Acknowledgments

This work was supported by the Important Scientific and Technical Achievements Transformation Project under Grant No. Z17-5-078, the Agricultural Science and Technology Innovation Team of the Science and Technology Department of Liaoning Province, the Large-scale Instrument Equipment Sharing Service Platform Capacity Building Fund under Grant No. kjhx2017028, the High-level Innovation Team Foreign Training Project under Grant No. 2018LNGXGJWPY-YB006, the Shenyang Science and Technology Plan Project under Grant Nos. F16-205-1-51, 17-65-7-00 and 17-231-1-04, and the Excellent Chinese and Foreign Youth Exchange Plant Project from the China Association for Science and Technology under Grant No. 2018CASTQNJL50. This project was supported by the Engineering Laboratory for Molecular Simulation and Designing of Drug Molecules of Liaoning and the Research Center for Computer Simulating and Information Processing of Bio-macromolecules of Liaoning Province.

Authors' Contributions

TZY and LZ are co–first authors and contributed equally to this work. TZY prepared the first draft of the paper and performed the primary computations for the analysis. LZ developed the main R program. LWY built the prediction website. FHW and SML performed the literature search and plotted the figures. LZ, HSL, and the other authors provided overall guidance, reviewed the results, or reviewed and contributed to this manuscript.

Conflicts of Interest

None declared.

References

1. World Health Organization. Classification of diabetes mellitus. Geneva: World Health Organization; Apr 21, 2019.
2. Carstensen B, Jørgensen ME, Friis S. The Epidemiology of Diabetes and Cancer. Curr Diab Rep 2014 Aug 26;14(10). [doi: 10.1007/s11892-014-0535-8] [Medline: 25156543]
3. Lu F, Lin K, Kuo H. Diabetes and the risk of multi-system aging phenotypes: a systematic review and meta-analysis. PLoS One 2009;4(1):e4144 [FREE Full text] [doi: 10.1371/journal.pone.0004144] [Medline: 19127292]
4. Wong E, Backholer K, Gearon E, Harding J, Freak-Poli R, Stevenson C, et al. Diabetes and risk of physical disability in adults: a systematic review and meta-analysis. Lancet Diabetes Endocrinol 2013 Oct;1(2):106-114 [FREE Full text] [doi: 10.1016/S2213-8587(13)70046-9] [Medline: 24622316]
5. Jeon CY, Murray MB. Diabetes Mellitus Increases the Risk of Active Tuberculosis: A Systematic Review of 13 Observational Studies. PLoS Med 2008 Jul 15;5(7):e152. [doi: 10.1371/journal.pmed.0050152]
6. Riza AL, Pearson F, Ugarte-Gil C, Alisjahbana B, van de Vijver S, Panduru NM, et al. Clinical management of concurrent diabetes and tuberculosis and the implications for patient services. Lancet Diabetes Endocrinol 2014 Sep;2(9):740-753 [FREE Full text] [doi: 10.1016/S2213-8587(14)70110-X] [Medline: 25194887]
7. Roy T, Lloyd CE. Epidemiology of depression and diabetes: a systematic review. J Affect Disord 2012 Oct;142 Suppl:S8-21. [doi: 10.1016/S0165-0327(12)70004-6] [Medline: 23062861]
8. Jacobs E, Hoyer A, Brinks R, Icks A, Kuß O, Rathmann W. Healthcare costs of Type 2 diabetes in Germany. Diabet Med 2017 Jun;34(6):855-861. [doi: 10.1111/dme.13336] [Medline: 28199029]
9. World Health Organization. Global Report On Diabetes. Geneva: World Health Organization; 2016.
10. Rawshani A, Rawshani A, Franzén S, Eliasson B, Svensson A, Miftaraj M, et al. Mortality and Cardiovascular Disease in Type 1 and Type 2 Diabetes. N Engl J Med 2017 Apr 13;376(15):1407-1418. [doi: 10.1056/NEJMoa1608664] [Medline: 28402770]
11. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet 2016 Apr 09;387(10027):1513-1530 [FREE Full text] [doi: 10.1016/S0140-6736(16)00618-8] [Medline: 27061677]
12. International Diabetes Federation. IDF Diabetes Atlas, 8th edition. Brussels: International Diabetes Federation; 2017.
13. Rawshani A, Rawshani A, Franzén S, Sattar N, Eliasson B, Svensson A, et al. Risk Factors, Mortality, and Cardiovascular Outcomes in Patients with Type 2 Diabetes. N Engl J Med 2018 Aug 16;379(7):633-644. [doi: 10.1056/NEJMoa1800256] [Medline: 30110583]
14. Simmons RK, Echouffo-Tcheugui JB, Sharp SJ, Sargeant LA, Williams KM, Prevost AT, et al. Screening for type 2 diabetes and population mortality over 10 years (ADDITION-Cambridge): a cluster-randomised controlled trial. Lancet 2012 Nov 17;380(9855):1741-1748 [FREE Full text] [doi: 10.1016/S0140-6736(12)61422-6] [Medline: 23040422]
15. Simmons RK, Rahman M, Jakes RW, Yuyun MF, Niggebrugge AR, Hennings SH, et al. Effect of population screening for type 2 diabetes on mortality: long-term follow-up of the Ely cohort. Diabetologia 2011 Feb;54(2):312-319. [doi: 10.1007/s00125-010-1949-8] [Medline: 20978739]
16. Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J Biomed Health Inform 2015 Mar;19(2):728-734. [doi: 10.1109/JBHI.2014.2325615] [Medline: 24860043]
17. Maniruzzaman M, Kumar N, Menhazul Abedin M, Shaykhul Islam M, Suri HS, El-Baz AS, et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput Methods Programs Biomed 2017 Dec;152:23-34. [doi: 10.1016/j.cmpb.2017.09.004] [Medline: 29054258]
18. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, Suri HS, Abedin MM, El-Baz A, et al. Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. J Med Syst 2018 Apr 10;42(5):92 [FREE Full text] [doi: 10.1007/s10916-018-0940-7] [Medline: 29637403]
19. 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 Dec;97:120-127 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.09.014] [Medline: 27919371]
20. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. Front Genet 2018;9:515 [FREE Full text] [doi: 10.3389/fgene.2018.00515] [Medline: 30459809]
21. Chung JW, Kim WJ, Choi SB, Park JS, Kim DW. Screening for pre-diabetes using support vector machine model. Conf Proc IEEE Eng Med Biol Soc 2014;2014:2472-2475. [doi: 10.1109/EMBC.2014.6944123] [Medline: 25570491]
22. Kuhn M. CRAN - R Project. 2019 Apr 18. caret: Classification and Regression Training URL: https://cran.r-project.org/web/packages/caret/ [accessed 2019-04-24]
23. Breiman L, Cutler A, Wiener M, Liaw A. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. 2018 Mar 25. URL: https://cran.r-project.org/web/packages/randomForest/ [accessed 2019-04-24]
24. Xavier R, Natacha T, Alexandre H, Natalia T, Frédérique L, Jean-charles S, et al. pROC:Display and Analyze ROC Curves. 2019 Mar 12. URL: https://cran.r-project.org/web/packages/pROC/index.html [accessed 2019-04-24]

25. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly), TU Wien. 2019 Mar 19. URL: https://cran.r-project.org/web/packages/e1071/index.html [accessed 2019-04-24]
26. Pozzolo A, Caelen O, Bontempi G. unbalanced: Racing for Unbalanced Methods Selection. 2015 Jun 26. URL: https://cran.r-project.org/web/packages/unbalanced/index.html [accessed 2019-04-24]
27. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2018 May 15;71(19):2199-2269 [FREE Full text] [doi: 10.1016/j.jacc.2017.11.005] [Medline: 29146533]
28. Mueller NT, Pereira MA. Leg length and type 2 diabetes: what's the link? Curr Opin Clin Nutr Metab Care 2015 Sep;18(5):452-456. [doi: 10.1097/MCO.0000000000000211] [Medline: 26167802]
29. Rapp K, Schroeder J, Klenk J, Ulmer H, Concin H, Diem G, et al. Fasting blood glucose and cancer risk in a cohort of more than 140,000 adults in Austria. Diabetologia 2006 May;49(5):945-952. [doi: 10.1007/s00125-006-0207-6] [Medline: 16557372]
30. Orsi E, Grancini V, Menini S, Aghemo A, Pugliese G. Hepatogenous diabetes: Is it time to separate it from type 2 diabetes? Liver Int 2017 Dec;37(7):950-962. [doi: 10.1111/liv.13337] [Medline: 27943508]
31. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967 Jun;54(1):167-179. [Medline: 6049533]
32. Cox DR. The Regression Analysis of Binary Sequences. J R Stat Soc B 2018 Dec 05;21(1):238-238. [doi: 10.1111/j.2517-6161.1959.tb00334.x]
33. Fisher R. The Use of Multiple Measurements in Taxonomic Problems. Ann Hum Genet 1936;7:179-188. [doi: 10.1111/j.1469-1809.1936.tb02137.x]
34. Gordon A, Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Biometrics 1984 Sep;40(3):874. [doi: 10.2307/2530946]
35. Tin KH. Random decision forests. 1995 Presented at: Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal; 1995; Quebec, Canada p. 278-282.
36. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
37. Liu X, Wu J, Zhou Z. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B Cybern 2009 Apr;39(2):539-550. [doi: 10.1109/TSMCB.2008.2007853] [Medline: 19095540]
38. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
39. Type 2 Diabetes Predicition Webset. URL: http://112.126.70.33/diabetes [accessed 2020-05-01]
40. Foreman KJ, Marquez N, Dolgert A, Fukutaki K, Fullman N, McGaughey M, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. Lancet 2018 Dec 10;392(10159):2052-2090 [FREE Full text] [doi: 10.1016/S0140-6736(18)31694-5] [Medline: 30340847]
41. Firouzi SA, Tucker LA, LeCheminant JD, Bailey BW. Sagittal Abdominal Diameter, Waist Circumference, and BMI as Predictors of Multiple Measures of Glucose Metabolism: An NHANES Investigation of US Adults. J Diabetes Res 2018 Jun;2018:3604108 [FREE Full text] [doi: 10.1155/2018/3604108] [Medline: 30018985]
42. Aune D, Ó HB, Vatten LJ. Resting heart rate and the risk of type 2 diabetes: A systematic review and dose--response meta-analysis of cohort studies. Nutr Metab Cardiovasc Dis 2015 Jun;25(6):526-534. [doi: 10.1016/j.numecd.2015.02.008] [Medline: 25891962]
43. Li X, Yu F, Zhou Y, He J. Association between alcohol consumption and the risk of incident type 2 diabetes: a systematic review and dose-response meta-analysis. Am J Clin Nutr 2016 Mar;103(3):818-829. [doi: 10.3945/ajcn.115.114389] [Medline: 26843157]
44. Valdez R, Yoon PW, Liu T, Khoury MJ. Family history and prevalence of diabetes in the U.S. population: the 6-year results from the National Health and Nutrition Examination Survey (1999-2004). Diabetes Care 2007 Oct;30(10):2517-2522. [doi: 10.2337/dc07-0720] [Medline: 17634276]
45. Feldman AL, Griffin SJ, Fhärm E, Norberg M, Wennberg P, Weinehall L, et al. Screening for type 2 diabetes: do screen-detected cases fare better? Diabetologia 2017 Nov;60(11):2200-2209 [FREE Full text] [doi: 10.1007/s00125-017-4402-4] [Medline: 28831538]
46. Lean ME, Leslie WS, Barnes AC, Brosnahan N, Thom G, McCombie L, et al. Primary care-led weight management for remission of type 2 diabetes (DiRECT): an open-label, cluster-randomised trial. Lancet 2018 Dec 10;391(10120):541-551. [doi: 10.1016/S0140-6736(17)33102-1] [Medline: 29221645]

Abbreviations

AUC: area under the curve

caret: Classification And REgression Training
NHANES: National Health and Nutrition Examination Survey
OR: odds ratio
PPV: positive predictive value
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 10.07.19; peer-reviewed by U Öberg, D Chao, A Van halteren, J Santos; comments to author 11.11.19; revised version received 22.12.19; accepted 07.02.20; published 18.06.20.

Please cite as:
Yang T, Zhang L, Yi L, Feng H, Li S, Chen H, Zhu J, Zhao J, Zeng Y, Liu H
Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation
JMIR Med Inform 2020;8(6):e15431
URL: https://medinform.jmir.org/2020/6/e15431
doi:10.2196/15431
PMID:32554386

Original Paper

Machine Learning–Based Signal Quality Evaluation of Single-Period Radial Artery Pulse Waves: Model Development and Validation

Xiaodong Ding[1*], BE; Feng Cheng[2*], PhD; Robert Morris[2], MA; Cong Chen[1], MM; Yiqin Wang[1], MD

[1]Shanghai Key Laboratory of Health Identification and Assessment, Laboratory of Traditional Chinese Medicine Four Diagnostic Information, Shanghai University of Traditional Chinese Medicine, Shanghai, China

[2]Department of Pharmaceutical Science, College of Pharmacy, University of South Florida, Tampa, FL, United States

[*]these authors contributed equally

Corresponding Author:
Yiqin Wang, MD
Shanghai Key Laboratory of Health Identification and Assessment
Laboratory of Traditional Chinese Medicine Four Diagnostic Information
Shanghai University of Traditional Chinese Medicine
1200 Cailun Road
Shanghai, 201203
China
Phone: 86 21 51322271
Email: wangyiqin2380@vip.sina.com

Abstract

Background: The radial artery pulse wave is a widely used physiological signal for disease diagnosis and personal health monitoring because it provides insight into the overall health of the heart and blood vessels. Periodic radial artery pulse signals are subsequently decomposed into single pulse wave periods (segments) for physiological parameter evaluations. However, abnormal periods frequently arise due to external interference, the inherent imperfections of current segmentation methods, and the quality of the pulse wave signals.

Objective: The objective of this paper was to develop a machine learning model to detect abnormal pulse periods in real clinical data.

Methods: Various machine learning models, such as k-nearest neighbor, logistic regression, and support vector machines, were applied to classify the normal and abnormal periods in 8561 segments extracted from the radial pulse waves of 390 outpatients. The recursive feature elimination method was used to simplify the classifier.

Results: It was found that a logistic regression model with only four input features can achieve a satisfactory result. The area under the receiver operating characteristic curve from the test set was 0.9920. In addition, these classifiers can be easily interpreted.

Conclusions: We expect that this model can be applied in smart sport watches and watchbands to accurately evaluate human health status.

(JMIR Med Inform 2020;8(6):e18134) doi:10.2196/18134

KEYWORDS

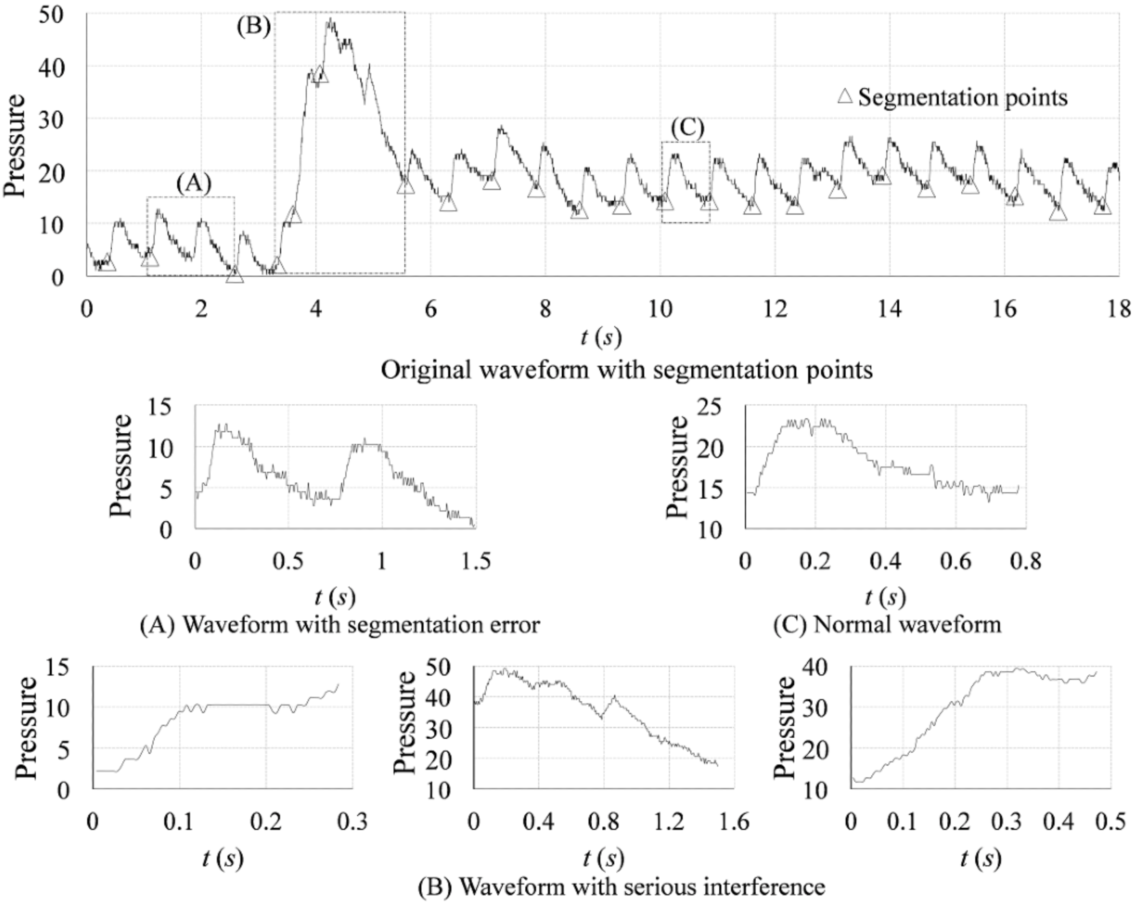
pulse wave; quality evaluation; single period; segmentation; machine learning

Introduction

Pulse-taking is widely used in disease diagnosis and personal health monitoring. For example, in traditional Chinese medicine (TCM), pulse-taking is an important approach to differentiate TCM syndrome patterns in which the physician uses their fingers to detect patients' pulsations. The radial artery is the most frequently used position for pulse-taking because the pulse wave of the radial artery contains abundant physiological information and is convenient for pulse-taking due to the accessibility of the vessels [1]. The development of smartwatches in recent years, coupled with pulse-taking analysis applications, enables individuals to monitor their pulse rates and physiological state throughout the day. As the number of smartwatch users expands, researchers are increasingly attempting to detect a variety of subclinical diseases such as atrial fibrillation (AF) by radial artery pulse waves in the early stages of disease progression. However, the majority of existing approaches are based on heart rate variability, which ignores important information contained in the changing pulse wave during a single cardiac cycle. The deep neural network models used for prediction are sometimes difficult to interpret [2]. Recently, researchers in the field of TCM diagnostics and hemodynamics have successfully utilized the information contained in single-period pulse waves not only to differentiate traditional syndrome patterns and diseases such as hypertension, diabetes, and other diseases not directly related to heart rhythm but also to fit modern clinical indices through objective recording [3-7]. Incorporating single-period pulse wave signals in smartwatches may improve the accuracy of existing applications in an interpretable way and expand the application scope of radial artery pulse waves.

In general, the radial artery pulse wave is a periodic signal. Therefore, it is necessary to segment the whole pulse wave series into periods before performing data mining of single-period pulse waves. However, the radial artery pulse wave signal is so weak that it is vulnerable to interference (such as breathing or vibration) during the acquisition process. These interferences can lead to waveform distortion, which increases the difficulty of segmenting the periods. In addition, no currently existing algorithm can extract single-period pulse wave signals from whole pulse wave series without error. Therefore, some pulse wave segments (abnormal segments) obtained by period segmentation are often remarkably different from the normal waveforms (Figure 1). These abnormal waveform pulses may affect future prediction results. As a result, automatically identifying these outliers from single-period pulse wave signals will significantly improve the accuracy of analysis.

Figure 1. Normal and abnormal pulse wave segments. A radial artery pulse wave series was segmented into periods by the segmentation method detailed in the Preprocessing section with α=.7. The segments of the original waveform between two adjacent segmentation points are regarded as single-period pulse waveforms. A and B show the abnormal segments caused by segmentation error and serious interference, respectively; C shows a normal segment; *t* (*s*): time in seconds.



Original waveform with segmentation points

(A) Waveform with segmentation error

(C) Normal waveform

(B) Waveform with serious interference

The early approach was to omit the waveform outliers that were too long or too short [8]. However, this method cannot be used to identify outliers with normal lengths. Thakker and Vyas [9] used dynamic time warping as a similarity measure in a pulse series in which the most dissimilar segments in the pulse series were classified as outliers. However, for patients with atrial fibrillation or other specific diseases, abnormal waveforms often appear in one series. In addition, significant outside interference may drastically reduce the number of normal segments. Due to these factors, it is difficult to discriminate between normal signals and outliers using this algorithm. This similarity is not the only important criterion for classifying pulse segments. Wang and Lu [10] utilized a k-nearest neighbor (KNN) classifier based on manual label data to measure the quality of the segmented single periods. However, the details and accuracy of the classifier were not shown. In a recent study, a method based on the Hilbert-Huang transform and an autoregressive moving average model was proposed to remove noise-induced mutations [11]. The accuracy of this method could reach 91.8% in a sample size of 207. However, because this method works on the entire pulse series, segmentation mistakes could not be identified. A consensus on the best method to evaluate the signal quality of a single-period pulse wave has not been reached. The purpose of this study was to utilize machine learning models to develop a signal quality evaluation model that can separate normal segments and abnormal segments. We expect that the model can be applied to help smart sport watches and watchbands evaluate human health status more accurately.

Methods

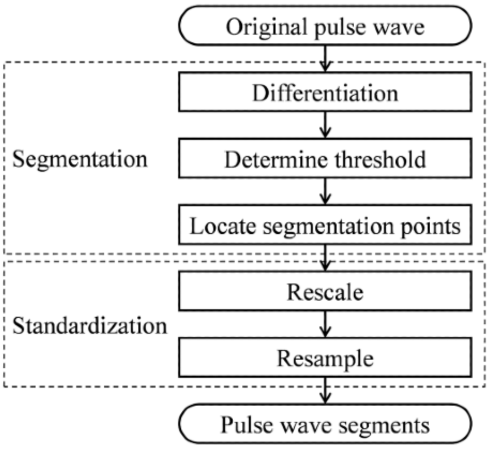
Data

In this study, the original radial artery pulse wave signals were taken from 390 outpatients at Shanghai Shuguang Hospital. All samples were split into an 80/20 ratio for training and testing sets. In other words, the data set was randomly divided into a training set with a capacity of 313 and a testing set with a capacity of 77.

Preprocessing

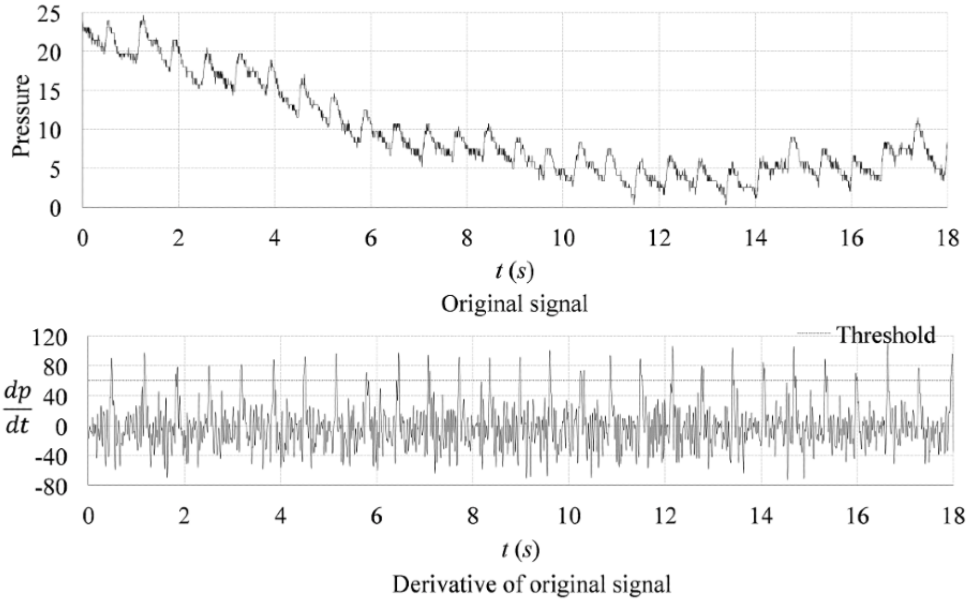
The steps of preprocessing, including segmentation and standardization, are illustrated in Figure 2.

Figure 2. The steps of data preprocessing. During segmentation, to reduce the influence of baseline wander, the derivative of the original signal was used to locate segmentation points by the threshold method. The corresponding segments of the original signal between two adjacent period segmentation points were single-period pulse wave segments. During standardization, each segment was rescaled and resampled to standardize its amplitude and length.



A simple segmentation method is the threshold method, which regards the minimum point below a specific threshold or the maximum point above a specific threshold as the period segmentation point. However, baseline wander is one of the most common forms of interference in pulse wave signals. Thus, it is difficult to apply the threshold method to the original signal. In contrast, as shown in Figure 3, the derivative of the original signal is almost entirely unaffected by baseline wander and also shows clearer segmentation points. Therefore, the threshold method can be used in the derivative for segmentation.

Figure 3. Sample pulse wave with baseline wander and the derivative of the pulse wave with an applied threshold. *t* (*s*): time in seconds.



To collect more normal and abnormal segments with different shapes, 5 different thresholds (between the maximum value of the derivative and 0) were selected for segmentation. If M is the maximum value of the derivative and α is a coefficient between 0 and 1, the threshold is given by

threshold = Mα (**1**)

In this study, for each patient's data, α takes 5 different values: 0.1, 0.3, 0.5, 0.7, and 0.9.

During segmentation, all threshold points of the derivative were first found. The first zero point of the derivative before each threshold point was defined as the period segmentation point, and the corresponding segments of the original signal between two adjacent period segmentation points were single-period waveforms. To avoid the presence of repetitive waveforms with significant similarities in the data set, only 5 segments were randomly selected for each threshold (if sufficient). No more than 25 segments were obtained from each patient by repeating the above process with different thresholds.

After segmentation, each segment was manually labeled as "normal" or "abnormal" by two expert annotators. A normal segment was required to have similar lengths, amplitudes, and shapes to most other segments in the same pulse series; be free of serious interference that could not be explained by the laws of physiology and pathology; have an approximately horizontal baseline, in which the difference between the values of the start point and the end point was not more than half the amplitude; and include only one complete cardiac cycle. The segments in which the two experts could not reach a consensus on their labels were not included in subsequent analyses. A total of 6832 segments were collected in the training set, of which 3974 (58.2%) were normal segments and 2858 (41.8%) were abnormal segments. In addition, a total of 1729 segments were collected in the testing set; 965 normal segments (55.8%) and 764 abnormal segments (44.2%) were identified.

The amplitudes and lengths of the segments differed from one another. To reduce the impact on the classification process, the original signals were standardized before classification. The segments were rescaled so that all values fell in the interval (0,1). If $X = \{x_1, x_2, \ldots, x_n\}$ was a segment, the rescaled segment Y was given by

Cubic spline interpolation was used to resample the segments to unify their lengths to a standard length. In this study, the trial values for the standard length were integers between 3 and 100. In general, the frequency range that contains useful information is below 25 hertz, and the cardiac cycle is no longer than 2 seconds. 100 sampling points are sufficient to contain all the useful information in one period. Hence, integers greater than 100 were not tested in this study.

Classification Methods

All 100 sampling points of the pulse wave segments were used as input features of the classifiers in this study. Three machine learning algorithms were applied to develop the classifiers:

1. KNN: The n_neighbors parameter was determined through cross-validation; the trial values for n_neighbors were integers in the range of 1-50.
2. Logistic regression: To reduce the influence of multicollinearity, L2 regularized logistic regression was chosen.
3. Support vector machine: support vector machine models with the radial basis function kernel (SVM-RBF), linear kernel (SVM-Linear), and 3-degree polynomial kernel (SVM-Poly) were applied. The cost and gamma parameters were determined through cross-validation, with trial cost values of 0.01, 0.1, 1, 10, and 100 and trial gamma values of 0.001, 0.01, 0.1, 1, and 10.

To estimate the accuracy of the models (out-of-sample accuracy), 10-fold cross-validation was applied. The most appropriate machine learning model was chosen by comparing the sensitivity, specificity and accuracy of the different algorithms. Based on the selected model, a reasonable value of standard input length was then identified by comparing the classification accuracies with different standard lengths.

In addition, the pulse waveforms that point at different positions in a cardiac cycle have different physiological significance and may influence the signal quality evaluation to varying degrees. To evaluate the contribution of each feature to the classification, recursive feature elimination was used to rank the features [12,13]. By excluding features one by one, it is possible to identify the smallest subset of features that can achieve satisfactory classification performance.

All the above steps were implemented in the training set. The final prediction model was validated on the testing set.

Results

Accuracy, Sensitivity, and Specificity of the Three Classification Models

The maximum accuracy, sensitivity, and specificity of each classification algorithm are presented in Table 1. All values are greater than 0.94, which indicates that all three algorithms performed similarly and effectively in waveform classification. To increase the simplicity and interpretability of the model, logistic regression was selected for further investigation.

Table 1. Comparison of the three classification algorithms, ±σ.

Method	Accuracy	Sensitivity	Specificity
KNN[a]	0.9732±0.0086	0.9901±0.0085	0.9494±0.0066
Logistic regression	0.9698±0.0122	0.9801±0.0141	0.9546±0.0128
Support vector machine			
SVM-RBF[b] (cost=10, gamma=0.1)	0.9797±0.0084	0.9862±0.0127	0.9691±0.0074
SVM-Linear[c] (cost=0.1, gamma=0.001)	0.9703±0.0124	0.9814±0.0137	0.9540±0.0140
SVM-Poly[d] (cost=1, gamma=0.1)	0.9756±0.0103	0.9866±0.0110	0.9594±0.0109

[a] KNN: k-nearest neighbors; n=6.

[b] SVM-RBF: support vector machine-radial basis function kernel.

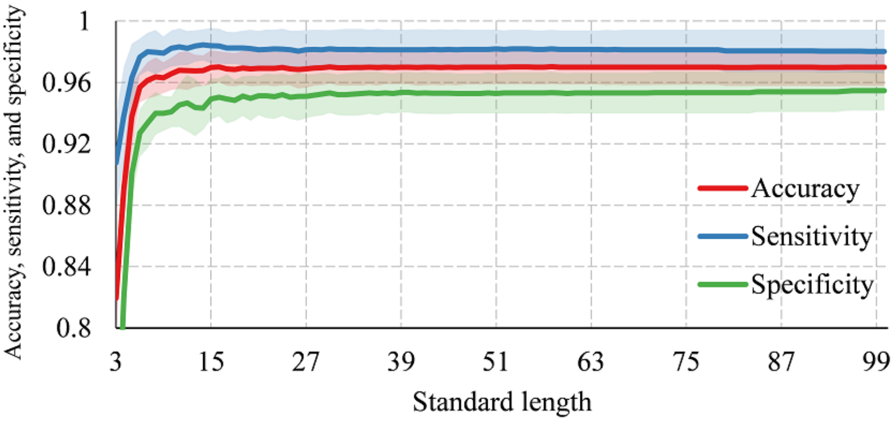
[c] SVM-Linear: support vector machine-linear kernel.

[d] SVM-Poly: support vector machine-3-degree polynomial kernel.

Standard Length

Figure 4 shows the performance of the logistic regression model with different standard lengths. When the standard length was >15, the model was stable and performed well; when the standard length was <15, the performance of the classifier deteriorated gradually as the standard length decreased. Therefore, 15 was selected as a reasonable value of the standard length for simplification of the model.

Figure 4. Accuracy, sensitivity, specificity, and standard deviation intervals of the classifier with different standard lengths.

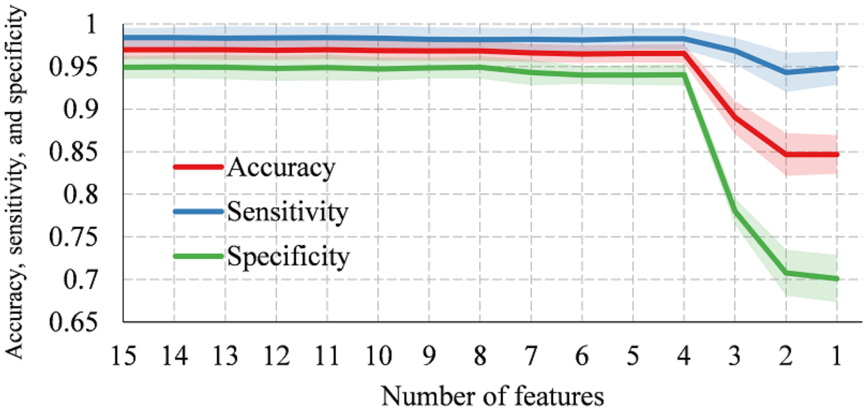


Each segment was resampled using the standard length of 15, and the importance ranking of the 15 feature points (Table 2) was then calculated based on the recursive feature elimination algorithm. When we sequentially eliminated the features one by one, as illustrated in Figure 5, the classifier performed equally well when the number of features was equal to or greater than 4. As shown in Table 2, the four most important feature points are the third, 14th, fourth, and first feature points. That is to say, a satisfactory logistic regression classification model can be established by using only four features (the third, 14th, fourth, and first points of the 15 feature points).

Table 2. Importance ranking of the feature points.

Importance ranking	Position of the feature point
1	3
2	14
3	4
4	1
5	7
6	15
7	13
8	11
9	6
10	9
11	10
12	2
13	8
14	5
15	12

Figure 5. Accuracy, sensitivity, and specificity of the classifier with sequentially eliminated input features and their standard deviation intervals.



Final Prediction Model

If x_n is the value of the nth feature point, and P is the probability that the segment is normal, the logistic regression classifier identified based on the training set can be given by

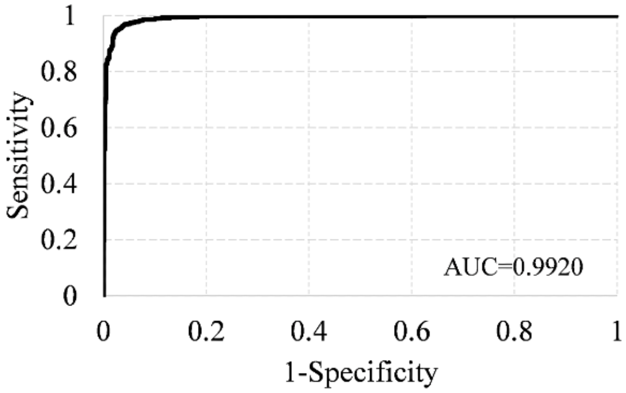
P = *sigmoid*(–9.6919x_1 + 8.2570x_3 + 8.9216x_4 – 7.9818x_{14} – 10.8732) **(3)**

where

The prediction model indicates that when x_1 and x_{14} are close to 0 and x_3 and x_4 are sufficiently large, the corresponding pulse wave segment can be classified as normal.

We applied this classifier to the testing set, and the receiver operating characteristic (ROC) curve is shown in Figure 6. The area under the curve (AUC) was 0.9920. Using the default threshold of 0.5, the accuracy, sensitivity, and specificity of the classifier were 0.9607, 0.9741, and 0.9437, respectively. This is consistent with the performance on the training set and achieves the expected result.

Figure 6. ROC curve of the final logistic regression classifier on the testing set. AUC: area under the curve.

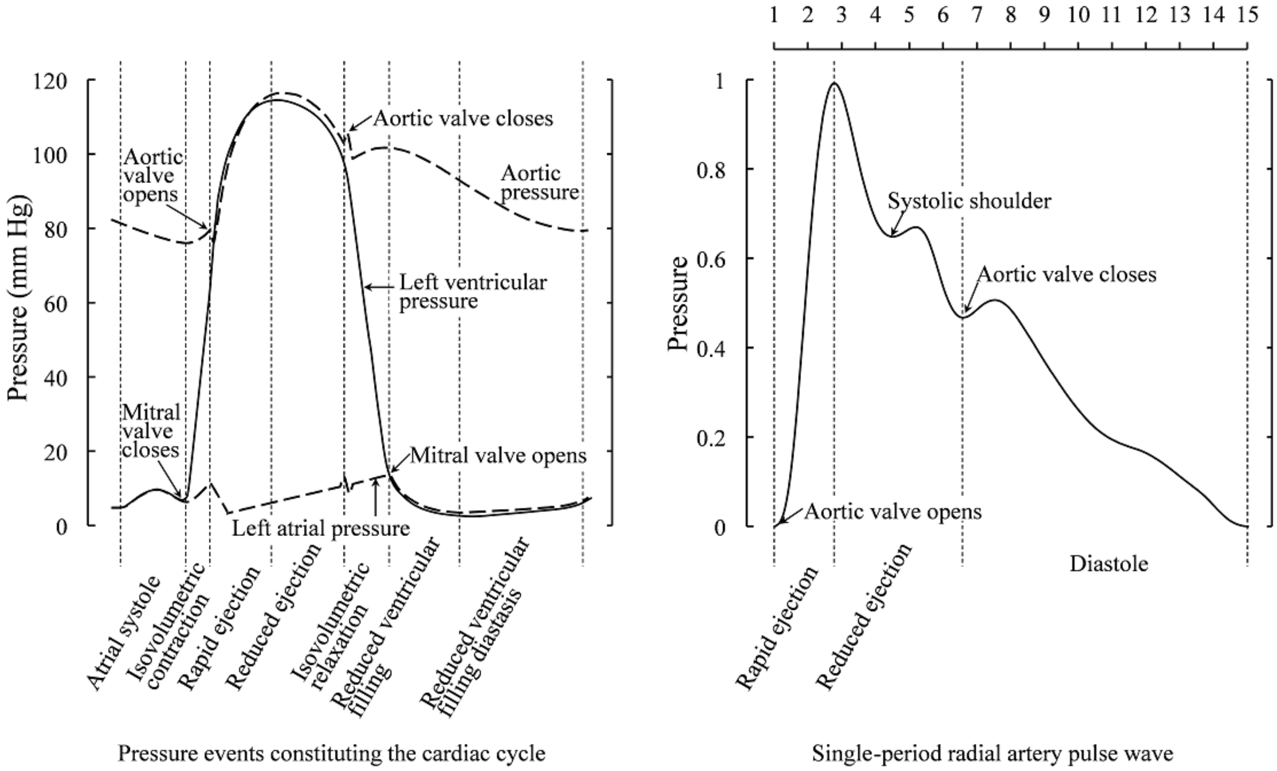


Discussion

In this paper, we compared the performance of three classification algorithms with different input features in the signal quality evaluation of single-period pulse waves. It was discovered that a logistic regression classifier with only four input features could achieve a satisfactory result. The equation of the final prediction model reveals that a pulse wave segment will be classified as normal only when x_1 and x_{14} are close to 0 and x_3 and x_4 are sufficiently large. This classifier is simple; however, it is consistent with the physiological process of the pulse wave.

A pulse wave is excited by cardiac ejection. As shown in Figure 7 [14,15], the fluctuation of the radial artery pulse wave corresponds to the events constituting the cardiac cycle. Therefore, a radial pulse wave is very similar to an aortic pulse wave. The pressure begins to rise rapidly as the aortic valve opens and ventricular ejection occurs. Shortly after ejection begins, the pressure reaches a peak and then gradually decreases. On the other side, the aortic pulse wave is greatly influenced by the reflection wave from the lower limbs, whereas the radial pulse wave is mainly influenced by the reflection wave from the upper limbs [16]. The peak of a radial pulse wave occurs much earlier than that of an aortic pulse wave due to differences in timing of the wave reflections in the upper limbs and the relatively distant lower body. When the reflection wave from the lower limbs and aortic valve closure propagates to the radial artery, the radial pulse wave correspondingly increases for a short time. These increases may not occur due to some physiological or pathological factors [17]. However, the rapid rise and gradual decline are essential features of a normal radial artery pulse wave.

Figure 7. Comparison of cardiac pressure and radial artery pressure in the cardiac cycle. The fluctuation of the radial artery pulse wave corresponds to the events constituting the cardiac cycle. For a normal waveform, both the starting and ending points should be close to 0, and the peak should appear near the third feature point. mm Hg: millimeters of mercury.



For a normal waveform, it is apparent that both the start and end points should be close to 0. A high start or end point indicates that the waveform exhibits significant interference. As a result of the relatively long duration of the diastole, the pulse wave remains low and steady at the end of the diastole. Both x_{14} and x_{15} are small in normal waveforms. However, as x_{15} is also the starting point of the next cardiac cycle, it is more likely to increase due to cardiac activity before systole and inaccurate cycle division. x_{14} is thus more representative of the end point than x_{15}. Correspondingly, at the beginning of the cardiac cycle, due to the rapid rise of the pulse wave in the systole, x_2 is not more representative than x_1. Therefore, x_1 and x_{14} were incorporated into the model to indicate the conditions of the starting and ending points. In addition, x_1 and x_{14} of an incomplete cycle will not be sufficiently small for the signal to be considered normal because the segmentation points of an incomplete cycle only contain signals of part of the cardiac cycle as opposed to both ends. Therefore, both x_1 and x_{14} can aid identification of the qualities of segmentation.

The peak of the radial artery pulse wave usually appears near x_3. Therefore, under normal conditions, x_3 should be close to 1. Due to the relatively slow rate of pressure drop during systole, x_4 does not have sufficient time to become very small. If either x_3 or x_4 is not sufficiently large, it is sufficient to prove that the waveform is abnormal. This can identify external interference; for example, the maximum value will not appear near x_3 due to the elevation of the latter part of the waveform, which can also lead to an anomaly in x_{14}. Furthermore, this change can aid the identification of segmentation errors. If we regard multiple cycles as one cycle, the first peak appears too early; as a result, x_3 and x_4 are not sufficiently large in most cases. If 5 or more cycles happen to be segmented together, one of x_3 or x_4 may be close to 1. However, under these circumstances, the waveforms rise and fall much more rapidly; thus, it is difficult for x_3 and x_4 to maintain large values simultaneously. Therefore, the incorporation of x_3 and x_4 into the model can help classify the waveform by detecting whether the peak of the waveform is located at the correct position.

In summary, the four input features of the logistic regression model are not only simple and effective but also interpretable. However, this model also has limitations. Most of our samples are free of heart disease, and the corresponding pathological characteristics will not be shown in the data. Consequently, this classifier only aimed to investigate common normal radial artery pulse waves. When searching for differences in the pulse waveforms of individuals without heart disease, we can effectively identify and eliminate abnormal segments with this classifier. However, for patients with some specific diseases, their pulse waves may change due to various pathological factors, which will ultimately lead to errors in classification. An example of this is sinus arrest, a condition in which the sinus node does not produce an impulse in one or more cardiac cycles; this causes the heartbeat to pause for a while, which will generate a long diastole segment in the pulse wave. This is a very important pathological signal; however, our model will classify it as a segmentation error due to the premature peak value. On the other hand, we may be able to use this property to improve the accuracy of some applications, such as distinguishing between premature contraction and atrial fibrillation. Premature contraction is the main cause of false positive error in AF detection algorithms [18]. There is still a certain proportion of normal segments in the pulse waves of patients with premature contraction, whereas AF will decrease the frequency of normal segments due to its irregular stroke volume and cardiac rhythm. This classifier may help distinguish the two cases by determining the ratio of normal segments to abnormal segments in the pulse wave series. In general, this classifier works well in normal cases, and its application scope can potentially expand according to its physiological significance. However, for some specific diseases, this classifier may lead to misclassification and even loss of key information. In the future, we hope to study the pulse wave characteristics of different diseases and distinguish them from random interference and the pulse wave characteristics of healthy people to subsequently improve the classifier and expand its application scope based on the new discoveries.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No. 81673880) and Shanghai Health Identification and Evaluation Laboratory Construction Funding (No. 13DZ2261000).

Authors' Contributions

XD, FC, and YW designed this study. XD and FC performed the data analysis. XD wrote the manuscript text, which was revised by FC and RM. CC helped with the sample collection. All authors read and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Wang Y. Diagnostics of Traditional Chinese Medicine. Beijing: Higher Education Press; Dec 2006.
2. Sajeev JK, Koshy AN, Teh AW. Wearable devices for cardiac arrhythmia detection: a new contender? Intern Med J 2019 May;49(5):570-573. [doi: 10.1111/imj.14274] [Medline: 31083804]

3. Huang C, Lin H, Liao W, Ceurvels W, Su S. Diagnosis of traditional Chinese medicine constitution by integrating indices of tongue, acoustic sound, and pulse. Eur J Integr Med 2019 Apr;27:114-120. [doi: 10.1016/j.eujim.2019.04.001]
4. Tsai Y, Chang Y, Huang Y, Jui-Shan Lin S, Lee S, Cheng Y, et al. The use of time-domain analysis on the choice of measurement location for pulse diagnosis research: A pilot study. J Chin Med Assoc 2019 Jan;82(1):78-85. [doi: 10.1016/j.jcma.2018.07.002] [Medline: 30839409]
5. Hao Y, Cheng F, Pham M, Rein H, Patel D, Fang Y, et al. A Noninvasive, Economical, and Instant-Result Method to Diagnose and Monitor Type 2 Diabetes Using Pulse Wave: Case-Control Study. JMIR Mhealth Uhealth 2019 Apr 23;7(4):e11959 [FREE Full text] [doi: 10.2196/11959] [Medline: 31012863]
6. Liu Z, Liu J, Wen B, He Q, Li Y, Miao F. Cuffless Blood Pressure Estimation Using Pressure Pulse Wave Signals. Sensors (Basel) 2018 Dec 02;18(12):4227 [FREE Full text] [doi: 10.3390/s18124227] [Medline: 30513838]
7. Poleszczuk J, Debowska M, Dabrowski W, Wojcik-Zaluska A, Zaluska W, Waniewski J. Patient-specific pulse wave propagation model identifies cardiovascular risk characteristics in hemodialysis patients. PLoS Comput Biol 2018 Sep 14;14(9):e1006417 [FREE Full text] [doi: 10.1371/journal.pcbi.1006417] [Medline: 30216341]
8. Xia C, Li Y, Yan J, Wang Y, Yan H, Guo R, et al. A practical approach to wrist pulse segmentation and single-period average waveform estimation. : IEEE; 2008 May 30 Presented at: 2008 International Conference on BioMedical Engineering and Informatics; 2008; Sanya, China. [doi: 10.1109/bmei.2008.140]
9. Thakker B, Vyas AL. Outlier pulse detection and feature extraction for wrist pulse analysis. International Journal of Biomedical and Biological Engineering 2009 Jan 01;3(7):127-130 [FREE Full text]
10. Wang D, Lu G. Period Segmentation for Wrist Pulse Signal Based on Adaptive Cascade Thresholding and Machine Learning. : IEEE; 2014 Presented at: 2014 International Conference on Medical Biometrics; 2014 May 30; Shenzhen, China. [doi: 10.1109/ICMB.2014.18]
11. Chen C. Research on detection method of abnormal pulse signal. Article in Chinese. Harbin, China: Harbin Institute of Technology; 2016. URL: https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CMFD&dbname=CMFD201801&filename=1017738492.nh [accessed 2020-02-05]
12. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46:389-422. [doi: 10.1023/a:1012487302797]
13. Zhou Q, Zhou H, Zhou Q, Yang F, Luo L. Structure damage detection based on random forest recursive feature elimination. Mech Syst Signal Pr 2014 May;46(1):82-90. [doi: 10.1016/j.ymssp.2013.12.013]
14. Fuster V, Harrington RA, Narula J, Eapen ZJ. Hurst's The Heart. New York: McGraw-Hill Education; 2017.
15. Fei Z, Zhang Z. Imaging and quantification of traditional pulse taking. Article in Chinese. Chinese Journal of Nature 1995;17(5):269-274.
16. Adji A, O'Rourke MF. Determination of central aortic systolic and pulse pressure from the radial artery pressure waveform. Blood Press Monit 2004 Jun;9(3):115-121. [doi: 10.1097/01.mbp.0000132426.32886.e0] [Medline: 15199304]
17. O'Rourke MF, Pauca A, Jiang X. Pulse wave analysis. Br J Clin Pharmacol 2001 Jun;51(6):507-522 [FREE Full text] [doi: 10.1046/j.0306-5251.2001.01400.x] [Medline: 11422010]
18. Bashar SK, Han D, Hajeb-Mohammadalipour S, Ding E, Whitcomb C, McManus DD, et al. Atrial Fibrillation Detection from Wrist Photoplethysmography Signals Using Smartwatches. Sci Rep 2019 Oct 21;9(1):15054 [FREE Full text] [doi: 10.1038/s41598-019-49092-2] [Medline: 31636284]

Abbreviations

AF: atrial fibrillation
AUC: area under the curve
KNN: k-nearest neighbor
ROC: receiver operating characteristic
SVM-Linear: support vector machine-linear kernel
SVM-Poly: support vector machine-polynomial kernel
SVM-RBF: support vector machine-radial basis function kernel
TCM: traditional Chinese medicine

Edited by G Eysenbach; submitted 05.02.20; peer-reviewed by V Osadchiy, G Stiglic; comments to author 16.03.20; revised version received 10.05.20; accepted 12.05.20; published 22.06.20.

Please cite as:
Ding X, Cheng F, Morris R, Chen C, Wang Y
Machine Learning–Based Signal Quality Evaluation of Single-Period Radial Artery Pulse Waves: Model Development and Validation
JMIR Med Inform 2020;8(6):e18134
URL: http://medinform.jmir.org/2020/6/e18134/
doi:10.2196/18134
PMID:32568091

Corrigenda and Addenda

Correction: Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier

Mark Singh[1*], BE(Elec), MD; Akansh Murthy[2*], BS; Shridhar Singh[3*]

[1]Carnegie Mellon University, University of Massachusetts Medical School, Braintree, MA, United States

[2]Massachusetts Institute of Technology, Cambridge, MA, United States

[3]Carnegie Mellon University, Pittsburgh, PA, United States

[*]all authors contributed equally

Corresponding Author:
Akansh Murthy, BS
Massachusetts Institute of Technology
77 Mass Ave
Cambridge, MA, 02139
United States
Phone: 1 6172531000
Email: ambshun@mit.edu

Related Article:

Correction of: https://medinform.jmir.org/2015/2/e17/

(JMIR Med Inform 2020;8(6):e21379) doi:10.2196/21379

In “Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier” (JMIR Med Inform 2015;3(2):e17) the authors noted an error in corresponding author’s details.

The corrected information is:

Akansh Murthy, BS
Massachusetts Institute of Technology
77 Mass Ave
Cambridge, MA, 02139
United States
Phone: 1 6172531000
Email: ambshun@mit.edu

The correction will appear in the online version of the paper on the JMIR Publications website on June 23, 2020, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories

Submitted 12.06.20; this is a non–peer-reviewed article;accepted 12.06.20; published 23.06.20.

Please cite as:
Singh M, Murthy A, Singh S
Correction: Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier
JMIR Med Inform 2020;8(6):e21379
URL: https://medinform.jmir.org/2020/6/e21379
doi:10.2196/21379
PMID:32574150

Original Paper

Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study

Christopher A Hane[1], PhD; Vijay S Nori[1], PhD; William H Crown[1], PhD; Darshak M Sanghavi[1], MD; Paul Bleicher[1], MD, PhD

OptumLabs, Optum, Cambridge, MA, United States

Corresponding Author:
Christopher A Hane, PhD
OptumLabs
Optum
1 Main St, 10th Floor
Cambridge, MA, 02142
United States
Phone: 1 6126326432
Email: christopher.hane@optum.com

Abstract

Background: Clinical trials need efficient tools to assist in recruiting patients at risk of Alzheimer disease and related dementias (ADRD). Early detection can also assist patients with financial planning for long-term care. Clinical notes are an important, underutilized source of information in machine learning models because of the cost of collection and complexity of analysis.

Objective: This study aimed to investigate the use of deidentified clinical notes from multiple hospital systems collected over 10 years to augment retrospective machine learning models of the risk of developing ADRD.

Methods: We used 2 years of data to predict the future outcome of ADRD onset. Clinical notes are provided in a deidentified format with specific terms and sentiments. Terms in clinical notes are embedded into a 100-dimensional vector space to identify clusters of related terms and abbreviations that differ across hospital systems and individual clinicians.

Results: When using clinical notes, the area under the curve (AUC) improved from 0.85 to 0.94, and positive predictive value (PPV) increased from 45.07% (25,245/56,018) to 68.32% (14,153/20,717) in the model at disease onset. Models with clinical notes improved in both AUC and PPV in years 3-6 when notes' volume was largest; results are mixed in years 7 and 8 with the smallest cohorts.

Conclusions: Although clinical notes helped in the short term, the presence of ADRD symptomatic terms years earlier than onset adds evidence to other studies that clinicians undercode diagnoses of ADRD. De-identified clinical notes increase the accuracy of risk models. Clinical notes collected across multiple hospital systems via natural language processing can be merged using postprocessing techniques to aid model accuracy.

(JMIR Med Inform 2020;8(6):e17819) doi:10.2196/17819

KEYWORDS

Alzheimer disease; dementia; health information systems; machine learning; natural language processing; health information interoperability

Introduction

Background

Worldwide, up to 77% of people with dementia are undiagnosed, and "lack of detection is a significant barrier to improving the lives of people with Alzheimer's disease and other dementias, their families and careers" [1]. This also implies that more than three-quarters of the patient population with dementia is not being referred for participation in clinical trials to study new potential treatments for neurodegenerative diseases. There are many factors influencing clinical trial recruitment for Alzheimer disease and related dementias (ADRD), including physician awareness of clinical trial opportunities, availability of study partners who can provide information on the study subject's functioning, the invasiveness of procedures often performed in Alzheimer trials, and concerns about labeling a patient with a serious dementia diagnosis with no known treatment [2].

Accurate prediction of the future onset of ADRD has several important practical applications. In particular, it facilitates the identification of individuals who are at high risk of developing ADRD to support the clinical development of novel treatments. Commonly, patients are identified after they are already symptomatic and have already experienced significant neurodegeneration. Screening patients into high-risk groups can facilitate the development of programs that investigate causal relations to specific ADRD etiologies and recruitment to clinical trials. Persons predicted to be at risk can also be offered the opportunity to plan more thoughtfully for the future while retaining their cognitive function.

A number of previous dementia risk models have been published in peer-reviewed literature [2-10]. Most of these studies used clinical data for model estimation, which limits their generalizability to other settings. This paper extends previous research by basing model estimation on a very large integrated dataset of medical claims and electronic health record (EHR) data as well as the use of more sophisticated machine learning estimation methods than those used in most previous studies. The use of medical claims and EHR data facilitates the use of the model in settings where large numbers of patients are treated, resulting in the identification of much larger potential patient populations for clinical trial recruitment [3-12].

Objectives

Nori et al [12] showed that machine learning models predict the onset of ADRD using medical claims and structured clinical data can have good performance near the time of onset and that performance diminishes with increasing time before onset. This study adds clinical notes data to those datasets to enhance the accuracy of the models and determines the prevalence of cognitive concerns in patient clinical notes up to 10 years before onset.

The quality of the clinical notes' models depends on common semantics in electronic medical record (EMR) systems. In their groundbreaking work on using EMR data for machine learning, Rajkomar et al [13] admitted, "Our current approach does not harmonize data between sites," but it can achieve similar accuracy at sites with sufficient volumes of data. Our study uses a dataset gathered from dozens of provider groups, mostly large integrated delivery network or hospital systems [14], and applies natural language processing (NLP) tools in a simple way to map semantically similar terms into concepts used in the models [15,16]. The processing of the clinical notes in this study favors automation, not clinical insight and expertise. This focus allows the methods to scale with little clinical intervention as new provider groups, and even new concepts are added to the data.

The use of a commercially available deidentified dataset will allow new studies to further refine the methods introduced here.

Methods

Overview

This study used deidentified medical claims and EHR data between 2007 and 2017 from the OptumLabs Data Warehouse (OLDW) [14]. The database contains longitudinal health information on enrollees and patients, representing a diverse mixture of ages, ethnicities, and geographical regions across the United States. The data in OLDW include medical and pharmacy claims, laboratory results, and enrollment records for commercial and Medicare Advantage enrollees. Clinical notes are available from a subset of EMR systems that chose to share these data. As this study involved analysis of preexisting, deidentified data, it was exempt from institutional review board approval [14].

Data Sets

This study uses and extends the clinical dataset of Nori et al [12]. That work created a matched case-control cohort of patients with onset of ADRD (cases) and patients with no history of any ADRD (controls). Index dates vary from 2009 to 2017 with 2 years of data per patient. In that earlier work, 7 different models with lead times of 0, 3, 4, 5, 6, 7, and 8 years to index were created from structured EHR and medical claims data to understand how predictive accuracy can be sustained over time. These models are called structured models because they only use structured data—diagnosis codes, procedure codes, and prescriptions—from the EMR and medical claims systems.

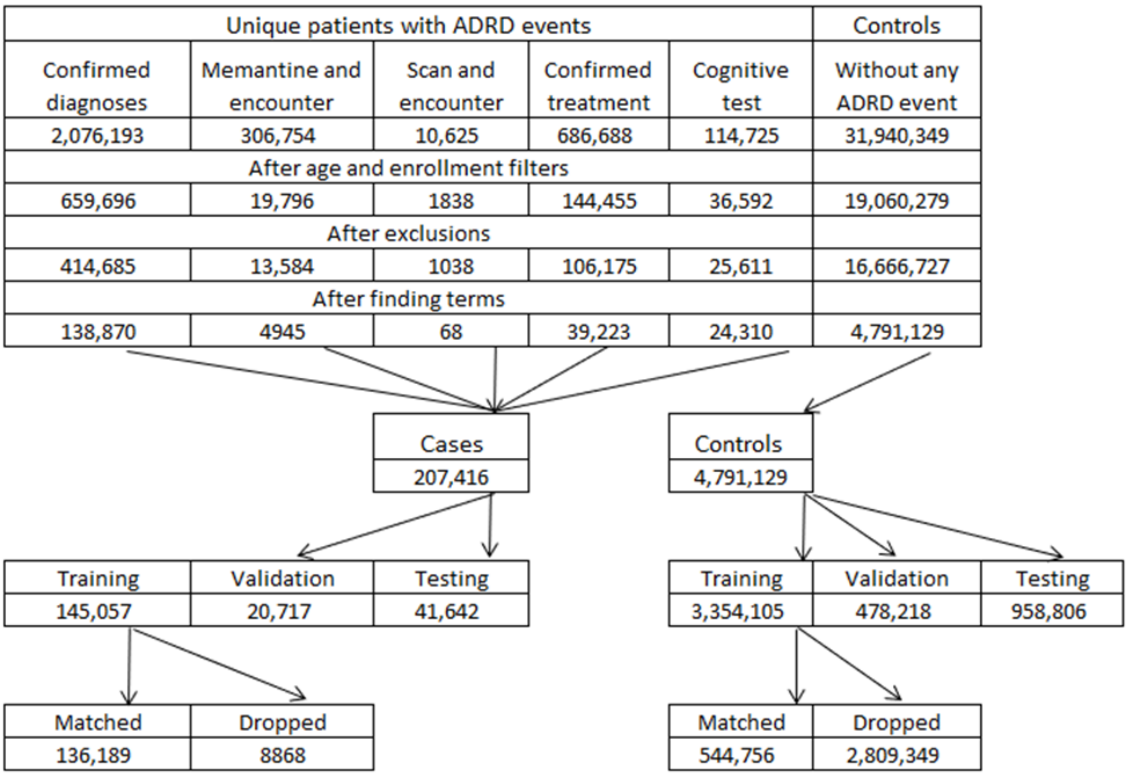
The outcome variable in this study was a confirmed incident diagnosis of ADRD, which includes mild cognitive impairment and forms of dementia but not alcohol-induced dementia [12]. These multiple forms of dementia diagnoses were included in the outcome after consultation with clinicians, and a review of the data indicated that specific diagnoses of a single type of dementia are less reliable, and elderly patients often have multiple dementias at onset [10,11].

This study uses the clinical notes of the same patients from structured data. Not all EMR systems provide raw clinical notes to the data collection process, so clinical notes are available because of data use agreements. Hence, patients' clinical notes data are missing due to legal agreements, not at random per patient and per encounter. To participate in the clinical notes' models, a patient must have 2 unique dates with a clinical note at least 31 days apart in the 2-year data collection period. The numbers of patients that met this threshold are provided in Table 1. No other adjustment for missing data was made. The attrition table of the population is shown in Figure 1. The first 3 filters are the same as those of Nori et al's study [12], with only the last filter of availability of clinical note data being specific to this analysis.

Table 1. Demographics of the study population.

Years to index date	Training set	N	Age, mean (SD)	Encounters, mean (SD)	Cases, n (%)	Females, n (%)
0	Matched training	680,945	74.3 (10.5)	30.5 (28.2)	136,189 (20.00)	417,390 (61.30)
3	Matched training	197,430	71.4 (10.2)	24.4 (21.9)	39,486 (20.00)	121,015 (61.30)
4	Matched training	130,270	70.4 (10.0)	22.5 (20.3)	26,054 (20.00)	79,795 (61.25)
5	Matched training	82,105	69.5 (9.8)	20.4 (18.8)	16,421 (20.00)	50,300 (61.26)
6	Matched training	47,555	68.6 (9.5)	18.3 (16.8)	9511 (20.00)	29,620 (62.29)
7	Matched training	23,455	67.6 (9.3)	16.6 (16.1)	4691 (20.00)	14,630 (62.37)
8	Matched training	7870	66.6 (9.1)	16.2 (15.8)	1574 (20.00)	4885 (62.07)
0	Validation	498,935	62.4 (11.3)	21.0 (21.8)	20,717 (4.15)	292,683 (58.66)
3	Validation	98,890	62.1 (10.8)	18.5 (17.9)	6525 (6.60)	60,560 (61.24)
4	Validation	61,471	61.7 (10.6)	17.4 (17.0)	4362 (7.10)	37,909 (61.67)
5	Validation	36,316	61.3 (10.4)	16.2 (15.9)	2763 (7.61)	22,464 (61.86)
6	Validation	18,888	61.0 (10.2)	15.2 (15.0)	1604 (8.49)	11,776 (62.35)
7	Validation	8186	60.6 (10.0)	14.2 (14.4)	806 (9.85)	5122 (62.57)
8	Validation	2660	60.0 (9.8)	14.0 (13.7)	292 (10.98)	1694 (63.68)
0	Test	1,000,448	62.3 (11.3)	21.0 (21.7)	41,642 (4.16)	587,326 (58.71)
3	Test	198,074	62.1 (10.8)	18.5 (18.0)	13,064 (6.60)	122,003 (61.59)
4	Test	122,939	61.7 (10.6)	17.4 (16.9)	8810 (7.17)	75,996 (61.82)
5	Test	72,304	61.3 (10.5)	16.1 (15.7)	5561 (7.69)	44,926 (62.13)
6	Test	37,938	61.0 (10.3)	14.9 (14.3)	3248 (8.56)	23,709 (62.49)
7	Test	16,487	60.8 (10.2)	14.1 (13.5)	1673 (10.15)	10,465 (63.47)
8	Test	5382	60.5 (10.1)	13.9 (13.8)	586 (10.89)	3474 (64.55)

Figure 1. Attrition table. ADRD: Alzheimer disease and related dementias.



Unique patients with ADRD events					Controls
Confirmed diagnoses	Memantine and encounter	Scan and encounter	Confirmed treatment	Cognitive test	Without any ADRD event
2,076,193	306,754	10,625	686,688	114,725	31,940,349
After age and enrollment filters					
659,696	19,796	1838	144,455	36,592	19,060,279
After exclusions					
414,685	13,584	1038	106,175	25,611	16,666,727
After finding terms					
138,870	4945	68	39,223	24,310	4,791,129

Figure 1 shows that patients who entered the cohort via a scan and a confirmatory diagnosis rarely had clinical notes that met the thresholds for inclusion (68 of 1038 patients). The mean number of encounters per patient is 21, but the patients that enter by the scan rule have only 4.4 encounters. This low encounter count before filtering by clinical note days indicates that there is little opportunity to have 2 days with clinical notes separated by a month. It is likely that these patients have encounters in a specialty setting where there is no complete view of the patient's health history.

As in Nori et al's study [12], cases and controls were matched for age, gender, number of encounters, and index year at a 1:4 ratio to reduce confounding of variables. This step is important to improve interpretability of variables and reduce multicollinearity because of age, which, if not performed, would lead to erroneous importance of age-related variables [17-21]. Due to filtering by days with a clinical note, the cases must be matched to controls anew in this work versus reusing the same sampling as in Nori et al's study [12].

Clinical Notes

The raw clinical notes go through the Optum proprietary NLP for determination of all patients' medical concept extraction. NLP concepts are identified and created based on broad topics such as medications, signs, disease and symptoms, measurements, and observations. The data are harvested from the clinical notes fields within the EMRs provided to Optum from over 50 large health care systems throughout the United States. The data used for the development of each NLP concept are deidentified, so the authors have no access to the raw notes.

The authors had access to the deidentified NLP data, which contains the date of the note, an occurrence date, a term, a sentiment, and possibly a family member. The terms are nouns, or abbreviations, extracted from the notes; sentiment describes the use of the noun (present, negative, possible, exhibit, exhibit.not, discuss, deny, concern, complain, etc). The content of the clinical notes are either about the patient or can be from a medical history where the content is about a family member. Family membership can be specific (mother, father, sibling, etc), vague (mother's relations, ancestor, and boyfriend), or a combination of relationships. The occurrence date may differ from the note date if the original text makes a temporal statement such as “a year ago the patient complained of…” This study's modeling uses the occurrence date of the term, if it exists, otherwise the date of the clinical note. We mapped the family members into 3 classes: immediate, family, and other (see Multimedia Appendix 1 for details). This mapping is based on wildcard word matching, so it is simple to implement, but may have errors. There are 29,528 unique terms and 1042 unique sentiments (42 positive sentiments such as “present,” “exhibit,” “observe”) with at least hundred patient clinical notes in all 7 yearly cohorts.

The NLP data have additional details, such as body location, severity, extent, and duration, that are not used in this study.

With the large number of unique sentiments (1042), the study decided to use only positive sentiment terms, indicating the presence of the term. In the raw notes, many negative terms are a collection of EMR survey questions *Patient denies smoking*, *Patient denies depression*, etc. These negative terms were excluded to reduce the complexity of processing the data and handling 1000 nonpositive sentiments.

Table 2 shows the highest 20 relative risk diagnoses in the onset year and their risks in the earlier years. The relative risk is the ratio of the probability of that diagnosis in cases versus controls. Each diagnosis must be supported by more than 10 cases, 10 controls, and 99 combined patients. Empty cells indicate that this threshold was not met. Years 7 and 8 had no unsuppressed values. These high-risk terms decay quickly over time; only 5 of the top 20 terms remain available for the model at year 3 and only 1 at year 4. Additional tables of terms, including tables with the most common terms, and the most common comorbidities are given in Multimedia Appendix 1.

Table 2. Top 20 relative risks of diagnosis.

Diagnosis	International Classification of Diseases, Ninth Revision code	Relative risk at years to index date				
		0	3	4	5	6
Wandering in diseases classified elsewhere	V403.1	21.57	—[a]	—	—	—
Unspecified senile psychotic condition	290.9	19.26	—	—	—	—
Unspecified persistent mental disorders due to conditions classified elsewhere	294.9	17.38	6.78	5.89	6.20	5.07
Senility without mention of psychosis	797.	16.48	—	—	—	—
Other general symptoms	780.9	16.27	—	—	—	—
Unspecified nonpsychotic mental disorder following organic brain damage	310.9	16.25	5.92	—	—	—
Other specified nonpsychotic mental disorders following organic brain damage	310.89	15.99	—	—	—	—
Other specified nonpsychotic mental disorder following organic brain damage	310.8	15.78	—	—	—	—
Other signs and symptoms involving cognition	799.59	15.06	4.45	—	—	—
Frontal lobe executive functional deficit	799.55	15.01	—	—	—	—
Dissociative amnesia	300.12	13.52	—	—	—	—
Personality change due to conditions classified elsewhere	310.1	13.52	4.42	—	—	—
Factitious disorder with predominantly psychological signs and symptoms	300.16	13.39	—	—	—	—
Psychotic disorder with delusions in conditions classified elsewhere	293.81	12.86	—	—	—	—
Confusional arousals	327.41	12.48	—	—	—	—
Visuospatial deficit	799.53	12.42	—	—	—	—
Reactive confusion	298.2	12.21	4.54	—	—	—
Subacute delirium	293.1	12.15	—	—	—	—
Alcohol-induced persisting amnestic disorder	291.1	12.07	—	—	—	—
Frontal lobe syndrome	310.0	12.07	—	—	—	—

[a]Not applicable.

Clinical Notes Clusters

With 29,528 unique terms in all the datasets, and without access to the algorithms that create the terms, the study needed to determine how the terms map to clinical concepts. In the raw clinical note, we expect that different clinicians will have alternative spellings—mi, ami, or acute myocardial infarction; htn or hypertension—depending on their training, the EMR they use, and many other factors. This study's upstream NLP does not map these terms into concepts but leaves them in their raw form. This creates a need to gather alternative spellings and related clinical terms into groups, or clusters, before using them. Without such a grouping, an individual term's impact may be diluted to the point of uselessness due to idiosyncratic abbreviations, spellings, and synonyms (eg, Alzheimer disease vs Alzheimer dementia). The methods here will ameliorate this situation. The terms are filtered to terms having at least 500 patients in any annual model, yielding 14,236 terms.

It would be possible, but time consuming, to map these terms to a medical ontology, but the study decided to pursue an algorithmic strategy relying on additional NLP processing. The end goal of this step is to map the terms to data-driven concepts that will group similar terms into more powerful machine learning features. The positive sentiment patient terms are processed into a sequence of terms, ordered by date, for each patient. Most of the clinical notes data lack a specific time of day, so the terms have no order other than a date. If the raw NLP provided a sequence number for the extracted terms, then that sequence could be used to order the patient's terms. Term sequences with less than 50 characters long are omitted. Due to database limits on the length of a single character field, the process needed to count characters in the concatenated terms. The choice was made to use these character counts to limit the patient stories used. In total, 50 characters is approximately 8 distinct terms. The word windows used are 10 words long.

At this step, there is a term-based *story* of 50-31,341 terms for each patient. The story file for all patients across all years is 5.9 gigabytes of text. The training text is collected and analyzed as 1 text file, with a row for each patient containing the patient's terms. To limit duplication of data from overlapping model years, model years 4 and 6 are not in the NLP training model (all year 4 terms are in either year 3 or year 5).

This text file can be analyzed using any NLP algorithm to build semantic knowledge among its words. The study chose to use Fasttext by the Facebook artificial intelligence research team for its speed and simplicity [22]. Fasttext builds a conditional probability model of term appearances in the context of their surrounding terms. The output of Fasttext is a numeric vector for each term. These term vectors are a meaningful mapping of each term to a vector space where the vector distance maintains word similarity. Thus, if 2 terms are very close to each other, measured by their vector distance, then they are synonyms. Alternate spellings of the same medical concept should be nearby in the vector space because the terms that surround them in the clinical notes will be similar. The study chose 100-dimension vectors to embed the terms.

To run the Fasttext algorithm, we chose the unsupervised continuous bag-of-words option with 8 epochs and a window of 10 terms. We explicitly turn off subwords because we do not want the algorithm associating the term alzheimers_disease with crohns_disease based on common subwords (syndrome is another confusing subword). Our reliance on terms from an upstream process means that misspellings are not an issue in this context. The unsupervised option means that Fasttext is finding semantic relations among the terms. The word window of 10 terms limits the probability model to overlapping sequences of 10 terms. The lack of sequence information on terms within a day means that the method needs more data to obtain a more accurate probability model of the text relations. Fasttext returns 11,061 terms due to its own filtering.

Once the study has the vector mapping of each term from Fasttext, the terms can be clustered into similar groups using the Euclidean distance of the terms as the similarity measure [23]. This is performed with the hclust function in R v3.5.1 [24]. As the goal of this clustering is to create features for the predictive model, we chose a large number of clusters, 1106 or 10% of the terms. A manual inspection of the clusters indicated that a much larger number of clusters (2212 or 5%) may split important sets of terms, and fewer clusters would merge groups that are less related to the outcome. Multimedia Appendix 1 shows the clusters for terms with individual memory and cognition terms. For example, the terms memory_loss, memory_issues, forgetful, memory, mild_cognitive_impairment, mci, recalling_issues, lewy_body_dementia, pseudodementia, memory_dysfunction, frontotemporal_dementia, and short_term_memory_loss all group together with a few more terms in 1 cluster. No manual editing of the clusters was performed. Note that the use of the embedding on training terms across all years, and the clustering of only those terms effectively omits novel terms present in the test data. However, because terms are the result of the upstream NLP over all years, the introduction of new terms in testing is rare. In production, new terms can be mapped into clusters if necessary.

The models use all the same medical features as the structured models and add 2 sets of features for the terms. One set of features counts the unique days with a specific term attributed to an individual or 1 of the 3 family types. The other set of features counts unique days at the clustered term level; thus, every term may appear twice, once specifically and once in its cluster.

Machine Learning

After computing these features, the same feature filtering method proposed by Nori et al [12] is applied. These filters remove features without sufficient support in the data as well as features whose ratio of matched to unmatched odds ratios are too extreme (see Multimedia Appendix 1). The filter based on the ratio of matched to unmatched odds avoids the inclusion of terms that are primarily associated with age and not with the outcome. Age-related terms can have a high unmatched odds ratio but a low matched odds ratio; thus, the ratio of these 2 values can filter these outliers. The ratio thresholds of 0.5 and 4 are used to remove features that are too skewed to age. For example, screening mammography has a very low unmatched odds ratio because it skews highly to younger women; the ratio of its unmatched to matched odds would be less than 0.5, and it would be removed from the model.

The models are fit with LightGBM [25], an algorithm that fits a gradient boosting machine to the 0 or 1 outcome variable using a series of decision trees. After solving the first model with a tree, another model is fit with a new outcome initialized to the residuals of the prior tree's fit. The series of residual fits optimize the loss function of the original model [26,27]. LightGBM also uses advanced methods such as sampling the feature space and pooling features into importance sets for improved performance. This study varied the parameters of LightGBM using a grid search; the model quality is assessed on the validation data, and the best model is selected. The study searched the feature fractions (.25, .2, .15), learning rates (.015, .01, .02), minimum data in leaf (1000, 800, 500), number of trees (300), and size of trees (127, 63). Each model was allowed to search independently for the parameter set that maximized its quality, as measured by the validation set's positive predictive value (PPV).

Results

This study reports 3 summary measures of model quality. The area under the curve (AUC) score is the probability that a random case score is higher than a random control. A drawback of AUC is that it is a global measure of discrimination, and it does not reflect the decision boundary to take action on a score from the model. With this in mind, we report 2 other measures, the PPV and lift.

PPV is the percentage of true positives in the at-risk population. To compute the PPV across this population, where the prevalence varies widely by age, we apply a threshold per age group where the number of patients at risk in each age group matches the age-based prevalence of cases.

Lift is the ratio of PPV to prevalence. Its values range from 0 to infinity. The lift reflects the improvement in the model over a random choice. For a rare disease, PPV may be low because the outcome is hard to detect and dividing by the prevalence provides a standardized way to correct for the prevalence.

Table 3 and Figure 2 show the model quality statistics for the baseline models without the clinical notes features and after adding the notes' terms and term clusters. In general, the models with clinical notes always have a higher AUC score, and the PPV and lift scores are higher in all but 1 year. Averaged over all models, the PPV was 5 points higher and the AUC was 4 points higher when terms and term clusters were included. Year 0 baseline values were reported by Nori et al [12].

Table 3. Quality of model fit on the test data.

Year	Sensitivity		Specificity		Area under the curve		Lift	
	Baseline	Clinical notes	Baseline	Clinical notes	Baseline	Clinical notes	Baseline	Clinical notes
0	0.45	0.68	0.98	0.99	0.84	0.94	13.92	16.39
3	0.27	0.30	0.95	0.95	0.67	0.70	4.12	4.62
4	0.27	0.29	0.94	0.95	0.66	0.69	3.80	4.03
5	0.25	0.28	0.94	0.94	0.61	0.68	3.23	3.60
6	0.25	0.24	0.93	0.93	0.62	0.63	2.91	2.84
7	0.24	0.26	0.91	0.92	0.62	0.68	2.39	2.52
8	0.25	0.26	0.91	0.91	0.59	0.58	2.34	2.43

Figure 2. Model quality measures. AUC: area under the curve; PPV: positive predictive value.

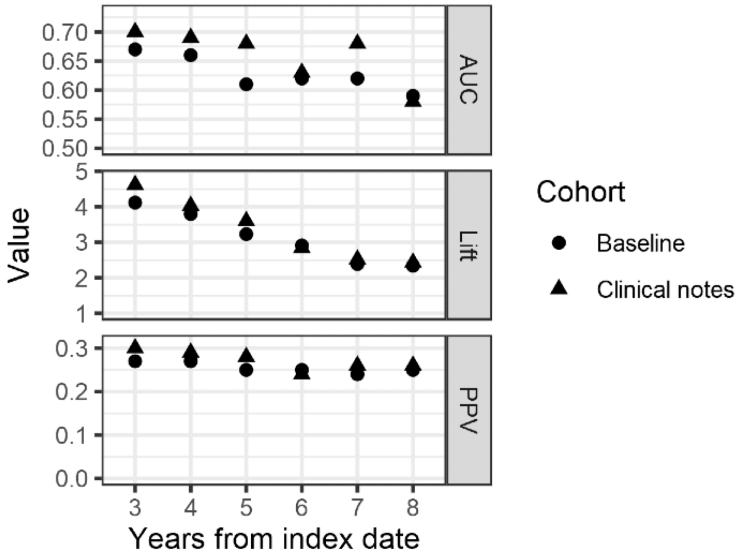


Table 4 shows the most important features in the onset year. The variable naming convention is *cls*, which is the prefix for cluster variables; *idv* for individual terms; *OFam* for other family; *IFam* for immediate family; *ETG* for episode treatment groups; *RXG* for medication therapeutic groups; *ICD* for International Classification of Disease version 9 (ICD-9); and *CPT* for Current Procedures and Terminology version 4. Tree-based machine learning algorithms rank variables by gain—how much the fit improves after that feature is used in a tree node. The gain is a dimensionless value, so we report the percentage of total gain attributed to each feature for all features up to 80% of the total gain. We refer to variables that meet this threshold as *important* variables. Important variables for the other years are shown in Multimedia Appendix 1.

Table 4. Important variables at onset (year 0) Total Gain (N) is 22,040,569.

Variable type	Variable name	Gain, n	Percent gain	Cumulative percent gain
cls	Dementia and Alzheimer dementia	3,298,549	15.0	15.0
cls	Memory loss and memory issues	2,833,536	12.9	27.8
idv	Dementia	2,162,843	9.8	37.6
idv	memory_issues	1,525,697	6.9	44.6
idv	memory_loss	1,498,113	6.8	51.4
idv	mild_cognitive_impairment	459,131	2.1	53.4
idv	Forgetful	419,780	1.9	55.3
cls	Alzheimer disease and other family memory issues	382,811	1.7	57.1
ETG	Neurological diseases signs and symptoms	378,955	1.7	58.8
idv	cognitive_impairment	346,991	1.6	60.4
idv	Memory	337,701	1.5	61.9
ICD	Altered mental status	275,533	1.3	63.2
idv	memory_lapses	256,076	1.2	64.3
idv	short_term_memory_loss	252,683	1.1	65.5
CPT	Neuropsychological testing (eg, Halstead-Reitan neuropsychological battery, Wechsler memory scales, and Wisconsin card sorting test), per hour of the psychologist's or physician's time, both face-to-face time administering tests to the patient and time interpreting these test results and preparing the report	245,279	1.1	66.6
cls	Cognitive impairment and hearing impairment	232,700	1.1	67.6
idv	Alzheimers_disease	221,324	1.0	68.6
cls	Cognitive issues and cognitive disorder	214,553	1.0	69.6
ETG	Mood disorder, depressed	214,171	1.0	70.6
CPT	Magnetic resonance (eg, proton) imaging, brain (including brain stem); without contrast material	213,180	1.0	71.5
ICD	Unspecified persistent mental disorders due to conditions classified elsewhere	174,643	0.8	72.3
cls	Memory lapses and concentrating	163,176	0.7	73.1
idv	getting_lost	159,014	0.7	73.8
CPT	Computed tomography, head or brain; without contrast material	150,658	0.7	74.5
cls	Family dementia and memory disturbance	125,350	0.6	75.1
ETG	Psychotic and schizophrenic disorders	121,595	0.6	75.6
dem	Age	118,598	0.5	76.1
RXG	Atypical antipsychotics	115,677	0.5	76.7
ETG	Mental disorders, organic and drug-induced	114,621	0.5	77.2
cls	Pain and tenderness	106,109	0.5	77.7
CPT	Neuropsychological testing (eg, Halstead-Reitan neuropsychological battery, Wechsler memory scales, and Wisconsin card sorting test), with qualified health care professional interpretation and report, administered by technician, per hour of technician time, face-to-face	100,425	0.5	78.1
dem	Number of encounters	94,671	0.4	78.6
RXG	Selective serotonin reuptake inhibitors	93,374	0.4	79.0
OFam	informant	92,567	0.4	79.4
idv	relaxing_issues	84,741	0.4	79.8
ICD	Depressive disorder, not elsewhere classified	77,831	0.4	80.1

Figures 3 and 4 extract terms that contain the phrases memory, dementia, and root *cognit* to see how the terms' prevalence varies over time. The top 15 terms ordered by matched case prevalence are displayed in 3 groups of 5 to allow scaling of the *y*-axis in each group. The plots show the prevalence of the unmatched validation data. The terms are present in each year if they are part of the model features for that year, but the filtering rules can omit them. For example, mild_cognitive_impairment appears in model years 0 through 6 but not in years 7 and 8 due to filtering. The cluster containing mild_cognitive_impairment is in all models and is important in all but the year 8 model. Figure 4 shows the plot of these terms in the control population, that is, a positive term for those without an incident diagnosis.

Figure 3. Frequency of cognitive terms in cases.

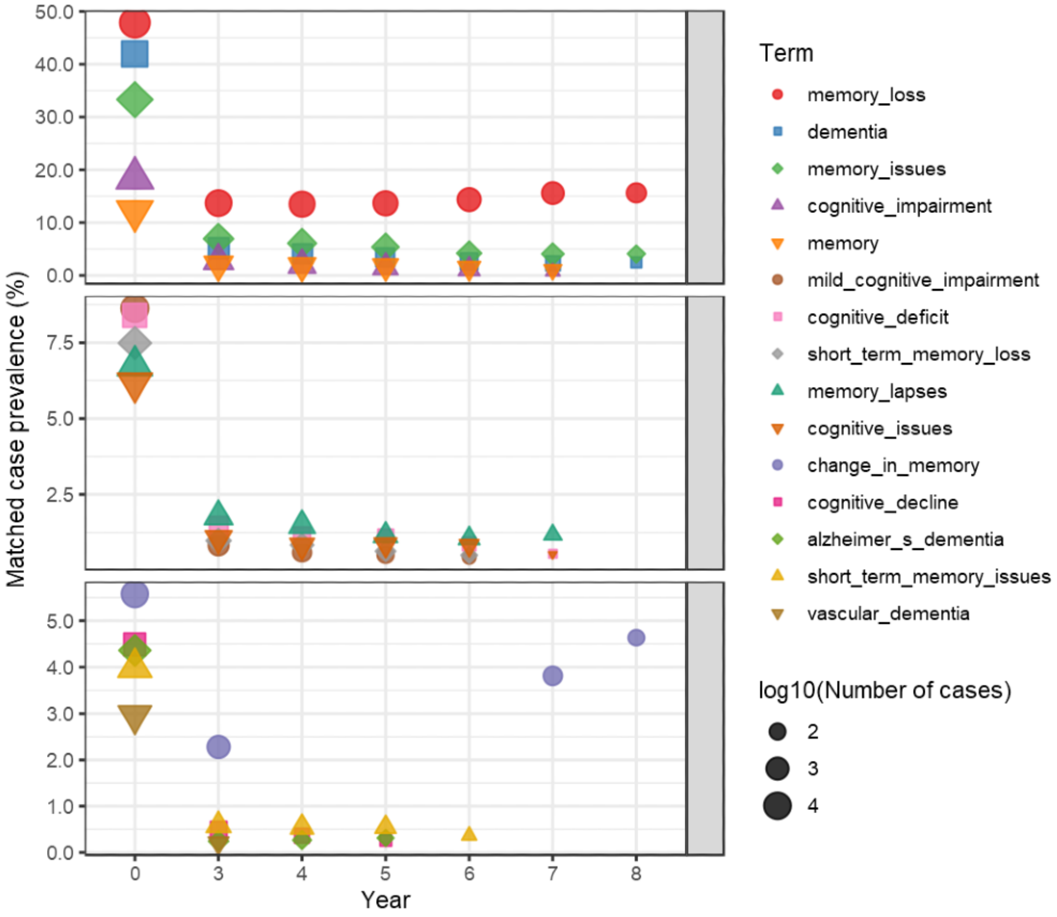
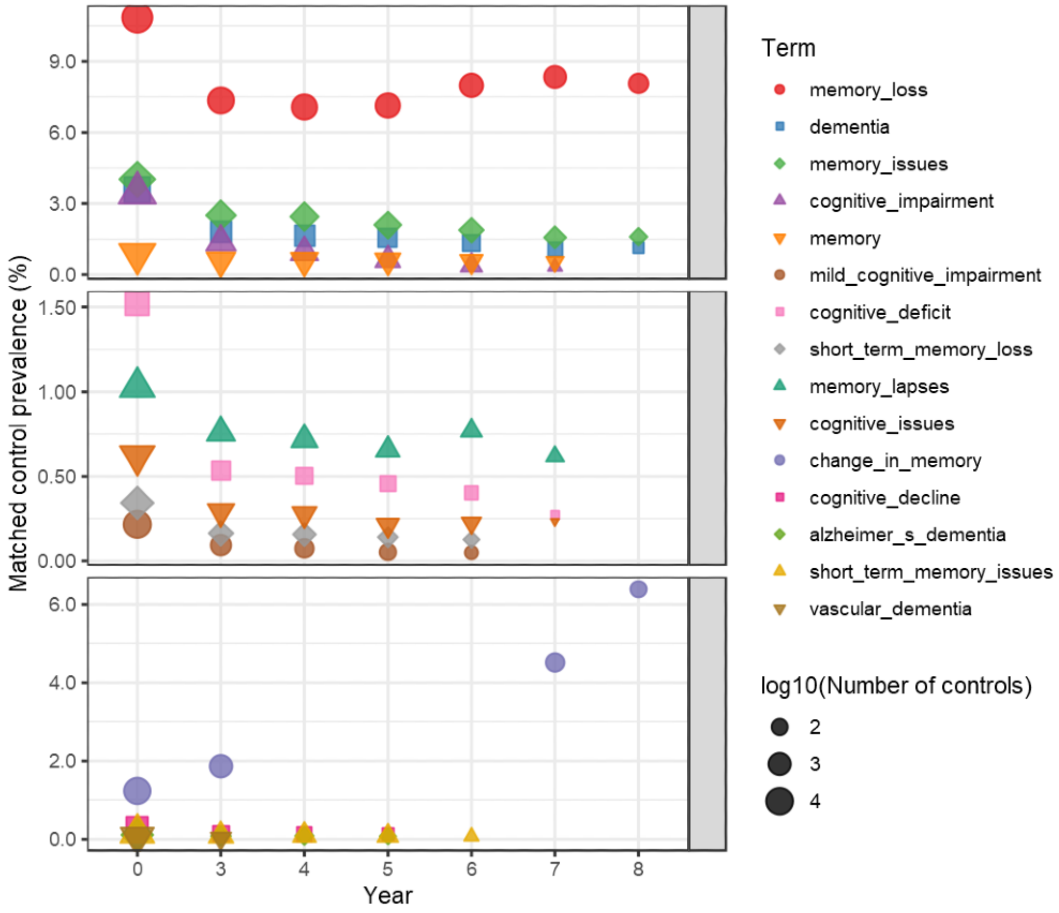


Figure 4. Frequency of cognitive terms in controls.



Discussion

Principal Findings

It is remarkable that even with all the diagnosis and prescription codes as well as neurological testing and radiology procedures available from the claims and structured EMR data, the clinical notes terms account for the first 8 and 21 of the 36 top predictors in the onset year model (Table 4). This indicates that the EHR data collection process collects important terms and that the NLP workflow is processing the clinical notes in a helpful manner. The most important non-note features are mood disorders, especially depression, psychoses, and prescription treatments for those disorders. It also may indicate that the clinical notes terms are a better indication of the prognostic symptoms than the structured data.

In the longer-term predictions, there are 3 data factors involved in decreasing accuracy. First, the cohorts rapidly decrease in size; for example, the training data at year 8 is 1% the size of the onset year (Table 1). This decrease in size is not just a survivorship issue, but the clinical notes data collection process was in its first year in year 8, so the diminishing size is a reflection of data collection growth from year 8 to year 0. Finally, there is commensurate growth in the features present in the model from 3450 to 7391 from year 8 to year 0, including all the medical coded features. As the scope of the data asset grows, it is possible that the future version of this model could perform better with little new modeling effort.

With ample evidence that ADRD is under coded [28-31], this dataset shows the existence of clinical notes with positive sentiment of many terms related to ADRD many years preceding the onset date of the cohort. The existence and ability of the model to extract terms such as memory loss, agitation, anxiety, and depression in the year 6 model (see Multimedia Appendix 1) demonstrates that these concerns are being coded in the clinical notes well before the ADRD diagnosis is recorded.

The family history terms are not as helpful as the individual terms. The only family term that survives in the important predictors is an immediate family history of Alzheimer disease only in the models for years 3, 4, and 5, but never over 0.3% of the total gain.

Figures 3 and 4 indicate that memory loss and other terms involving cognition and dementia are present at higher rates than one may expect. Memory loss is present in more than 13.52% (3522/26,054) of cases throughout the model years. Clearly, one wonders if the cases with these terms have a delayed diagnosis in the structural data. Furthermore, Figure 4 shows that the controls had at least 7.07% (7364/104,216) prevalence of memory loss in all model years. Although not all memory loss is an indication of ADRD, these prevalences are in a population whose mean age is in the low 60’s (Table 1) and could be an indication of under coding ADRD in the controls. However, the crux of the issue is that these memory terms in clinical notes may not reflect the underlying physiological changes of dementia or under coding.

Comparison With Previous Work

The structured models fit by Nori et al [12] found evidence of increased mental health, neurological testing, and anticholinergic risk factors found in other studies, as well as cardiovascular risk, which has been associated with vascular dementia [31-33]. The structured models in Nori et al's study [12] did not confirm diabetes mellitus as a risk factor, as found in Haan's study [33]. However, this study does find diagnoses of diabetes mellitus as important in the 3-, 4-, 5-, and 8-year models both as a coded diagnosis as well as in the notes. In addition, the clinical notes as far back as 6 years do identify metabolic syndrome and a cluster of terms related to insulin resistance as important (see Multimedia Appendix 1). This is important because there is an ICD code for metabolic syndrome, but that code does not surface in the structured data models. This provides some evidence in support of Haan's study [33], which is not present in the structured model.

Several previous studies have attempted to model the risk of Alzheimer and related disorders. Most of these have been small studies using detailed clinical data [3-9]. Recently, several studies have used widely available claims data to predict the onset of Alzheimer and related disorders [10-12]. These claims-based models achieved similar results similar to those of earlier studies (AUC ranging from 0.60 to 0.78). Our study shows that incorporating EHR data into the analysis results in significant additional improvements in the performance of models predicting Alzheimer and related disorders.

Limitations

This analysis was conducted using administrative claims and EHR data. The accuracy of diagnostic coding is a known issue with claims data. In the case of dementia, diagnostic inaccuracy is especially challenging in distinguishing between its different forms. In related work, we are exploring methods that explicitly address errors in the labeling of those who have dementia, and this appears to be promising [12]. Nevertheless, it is interesting to note that, despite the diagnostic coding issues, our models perform on par (or better) than previously published models using much deeper clinical information.

Conclusions

Our findings have important implications for the usefulness of predictive models based on administrative claims and EMR data to identify individuals at risk of dementia. Given the widespread availability of claims data that are already routinely used to identify individuals for interventions such as disease management programs, it is clear that predictive models could clearly be much more widely used to support individuals at risk of dementia in the community to help delay or even prevent institutionalization in nursing homes as well as aid in financial planning and provide other support needed by families having a member with dementia.

Similarly, the widespread availability of EHRs in clinical settings would enable clinicians to make use of predictive models to support their patients with dementia and their families. In the rarer settings where both claims and EHR data are available, our findings indicate that predictive models will be even more effective at identifying patients with dementia who could benefit from social support.

A second, very important, application of machine learning models is to identify patients for recruitment into clinical trials. Collecting the clinical data needed for screening dementia patients for clinical trials is extremely expensive—in the case of Alzheimer disease estimated to be over US $4000 to screen patients with cognitive assessments and positron emission tomography scans [34,35]. This cost is inflated by the need to screen many individuals for every individual identified. Any tool that reduces the number of patients needing to be screened will reduce the cost of patient recruitment. As reported in Table 3, models using EHR data correctly identified 2.5 times the number of patients with dementia relative to baseline prevalence 8 years to diagnosis and 16.4 times the baseline prevalence at the time of diagnosis. Although model performance declines further from diagnosis, these results suggest that predictive models based on machine learning methods could also be helpful in identifying patients earlier in their disease course. This is important for both the provision of social support and clinical trial recruitment. In the case of social support, it may well be the case that intervening earlier will be more effective in delaying nursing home institutionalization, and it would certainly give families more time to prepare for such an outcome. In the case of clinical trials, it is possible that recruiting patients who are earlier in the disease course may improve the effectiveness of pharmacologic interventions, which, to date, have been of little clinical value.

Clinical notes data extracted into deidentified structured tables can be useful in adding value to models built with structured data. The accuracy of the onset year model is much higher than that of other models in the literature (94% AUC vs in the 70% range) [8,20-33]. This ability to discriminate with a PPV of 68% (51% higher than without notes) means that this model can be an effective screening tool for patient and provider follow-up. The rapid decline in model quality beyond diagnosis limits the utility of the model for long-term prediction. It is unclear if this decline can be easily remedied once more clinical notes data are available or if this indicates a more important issue of primary data collection and under coding with more challenging remedies.

Clustering terms from the deidentified clinical notes helps to overcome variation in how clinical notes are written across diverse provider groups. The term clusters also boost the strength of the group by combining similar concepts into a coherent feature that improves prediction.

Future work should focus on semisupervised approaches that expand the data available for training by learning to label data in a consistent manner. Work on semisupervised methods can also help enhance the reliability of case/control labeling for ADRD, as started by Nori et al [12].

Recent work by Xie [36] shows promise for augmenting labeled data with use cases outside the medical domain. It is not yet clear if these methods can augment/perturb medical record data in a way that can boost model performance.

Acknowledgments

The authors thank Pamela Hansen of OptumLabs for her diligent support of the project. The project sponsorship from OptumLabs, Global CEO Initiative, Biogen, Janssen Pharmaceutical, and Merck was graciously received. A donation of computer hardware by Dell allowed the project to fit these large models efficiently.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Project analytical overview and additional results.
[DOCX File , 76 KB - medinform_v8i6e17819_app1.docx]

References

1. Alzheimer's Disease International. World Alzheimer Report 2011: The Benefits of Early Diagnosis and Intervention URL: https://www.alz.co.uk./research/world-report-2011, [accessed 2019-06-15]
2. Watson JL, Ryan L, Silverberg N, Cahan V, Bernard MA. Obstacles and opportunities in Alzheimer's clinical trial recruitment. Health Aff (Millwood) 2014 Apr;33(4):574-579 [FREE Full text] [doi: 10.1377/hlthaff.2013.1314] [Medline: 24711317]
3. Barnes DE, Beiser AS, Lee A, Langa KM, Koyama A, Preis SR, et al. Development and validation of a brief dementia screening indicator for primary care. Alzheimers Dement 2014 Nov;10(6):656-65.e1 [FREE Full text] [doi: 10.1016/j.jalz.2013.11.006] [Medline: 24491321]
4. Byeon H. A prediction model for mild cognitive impairment using random forests. Int J Adv Comput Sci App 2015;6(12):8. [doi: 10.14569/ijacsa.2015.061202]
5. Barnes D, Covinsky K, Whitmer R, Kuller L, Lopez O, Yaffe K. Predicting risk of dementia in older adults: the late-life dementia risk index. Neurology 2009 Jul 21;73(3):173-179 [FREE Full text] [doi: 10.1212/WNL.0b013e3181a81636] [Medline: 19439724]
6. Exalto LG, Biessels GJ, Karter AJ, Huang ES, Katon WJ, Minkoff JR, et al. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study. Lancet Diabetes Endocrinol 2013 Nov;1(3):183-190 [FREE Full text] [doi: 10.1016/S2213-8587(13)70048-2] [Medline: 24622366]
7. Exalto LG, Quesenberry CP, Barnes D, Kivipelto M, Biessels GJ, Whitmer RA. Midlife risk score for the prediction of dementia four decades later. Alzheimers Dement 2014 Sep;10(5):562-570 [FREE Full text] [doi: 10.1016/j.jalz.2013.05.1772] [Medline: 24035147]
8. Kivipelto M, Ngandu T, Laatikainen T, Winblad B, Soininen H, Tuomilehto J. Risk score for the prediction of dementia risk in 20 years among middle aged people: a longitudinal, population-based study. Lancet Neurol 2006 Sep;5(9):735-741. [doi: 10.1016/S1474-4422(06)70537-3] [Medline: 16914401]
9. Reitz C, Tang M, Schupf N, Manly JJ, Mayeux R, Luchsinger JA. A summary risk score for the prediction of Alzheimer disease in elderly persons. Arch Neurol 2010 Jul;67(7):835-841 [FREE Full text] [doi: 10.1001/archneurol.2010.136] [Medline: 20625090]
10. Albrecht JS, Hanna M, Kim D, Perfetto EM. Predicting diagnosis of Alzheimer's disease and related dementias using administrative claims. J Manag Care Spec Pharm 2018 Nov;24(11):1138-1145 [FREE Full text] [doi: 10.18553/jmcp.2018.24.11.1138] [Medline: 30362918]
11. Nori VS, Hane CA, Martin DC, Kravetz AD, Sanghavi DM. Identifying incident dementia by applying machine learning to a very large administrative claims dataset. PLoS One 2019;14(7):e0203246 [FREE Full text] [doi: 10.1371/journal.pone.0203246] [Medline: 31276468]
12. Nori V, Hane CA, Crown WH, Au R, Burke WJ, Sanghavi DM, et al. Machine learning models to predict onset of dementia: a label learning approach. Alzheimers Dement (N Y) 2019;5:918-925 [FREE Full text] [doi: 10.1016/j.trci.2019.10.006] [Medline: 31879701]
13. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
14. Wallace P, Shah N, Dennen T, Bleicher P, Bleicher PD, Crown WH. Optum labs: building a novel node in the learning health care system. Health Aff (Millwood) 2014 Jul;33(7):1187-1194. [doi: 10.1377/hlthaff.2014.0038] [Medline: 25006145]
15. Optum. 2019. Clinformatics Data Mart URL: https://www.optum.com/content/dam/optum/resources/productSheets/Clinformatics_for_Data_Mart.pdf [accessed 2019-07-15]
16. Gunaseelan V, Kenney B, Lee J, Hu H. Databases for surgical health services research: clinformatics data mart. Surgery 2019 Apr;165(4):669-671. [doi: 10.1016/j.surg.2018.02.002] [Medline: 29555196]
17. O’Brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007 Mar 13;41(5):673-690. [doi: 10.1007/s11135-006-9018-6]

18. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 2012 May 18;36(1):27-46. [doi: 10.1111/j.1600-0587.2012.07348.x]
19. Kraha A, Turner H, Nimon K, Zientek LR, Henson RK. Tools to support interpreting multiple regression in the face of multicollinearity. Front Psychol 2012;3:44 [FREE Full text] [doi: 10.3389/fpsyg.2012.00044] [Medline: 22457655]
20. Sandri M, Zuccolotto P. A bias correction algorithm for the gini variable importance measure in classification trees. J Comput Graph Stat 2008 Sep;17(3):611-628. [doi: 10.1198/106186008x344522]
21. Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008 Jul 11;9:307 [FREE Full text] [doi: 10.1186/1471-2105-9-307] [Medline: 18620558]
22. FastText. What is FastText? URL: https://fasttext.cc/index.html [accessed 2019-07-15]
23. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. ArXivCs 2016:- epub ahead of print - 1607.01759. [doi: 10.18653/v1/e17-2068]
24. The R Project for Statistical Computing. 2017. R: A Language and Environment for Statistical Computing URL: https://www.R-project.org/ [accessed 2017-01-06]
25. LightGBM's Documentation!-Read the Docs. Welcome to LightGBM's Documentation! URL: https://lightgbm.readthedocs.io/en/latest/ [accessed 2019-07-15]
26. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Statist 2001 Oct;29(5):1189-1232. [doi: 10.1214/aos/1013203451]
27. Breiman L. Prediction games and arcing algorithms. Neural Comput 1999 Oct 1;11(7):1493-1517. [doi: 10.1162/089976699300016106] [Medline: 10490934]
28. Chodosh J, Petitti D, Elliott M, Hays R, Crooks V, Reuben D, et al. Physician recognition of cognitive impairment: evaluating the need for improvement. J Am Geriatr Soc 2004 Jul;52(7):1051-1059. [doi: 10.1111/j.1532-5415.2004.52301.x] [Medline: 15209641]
29. Newcomer R, Clay T, Luxenberg J, Miller R. Misclassification and selection bias when identifying Alzheimer's disease solely from medicare claims records. J Am Geriatr Soc 1999 Feb;47(2):215-219. [doi: 10.1111/j.1532-5415.1999.tb04580.x] [Medline: 9988293]
30. Romero J, Benito-León J, Mitchell AJ, Trincado R, Bermejo-Pareja F. Under reporting of dementia deaths on death certificates using data from a population-based study (NEDICES). J Alzheimers Dis 2014;39(4):741-748. [doi: 10.3233/JAD-131622] [Medline: 24254704]
31. Adelborg K, Horváth-Puhó E, Ording A, Pedersen L, Sørensen HT, Henderson V. Heart failure and risk of dementia: a Danish nationwide population-based cohort study. Eur J Heart Fail 2017 Feb;19(2):253-260 [FREE Full text] [doi: 10.1002/ejhf.631] [Medline: 27612177]
32. Rusanen M, Kivipelto M, Levälahti E, Laatikainen T, Tuomilehto J, Soininen H, et al. Heart diseases and long-term risk of dementia and Alzheimer's disease: a population-based CAIDE study. J Alzheimers Dis 2014 Aug 11;42(1):183-191. [doi: 10.3233/jad-132363]
33. Haan M. Therapy insight: type 2 diabetes mellitus and the risk of late-onset Alzheimer's disease. Nat Clin Pract Neurol 2006 Mar;2(3):159-166. [doi: 10.1038/ncpneuro0124] [Medline: 16932542]
34. Boustani M, Callahan CM, Unverzagt FW, Austrom MG, Perkins AJ, Fultz BA, et al. Implementing a screening and diagnosis program for dementia in primary care. J Gen Intern Med 2005 Jul;20(7):572-577 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.0126.x] [Medline: 16050849]
35. Cruz JS. Today's Geriatric Medicine. 2016. Amyloid PET Imaging for Alzheimer's Disease Diagnosis URL: https://www.todaysgeriatricmedicine.com/news/ex_112116.shtml [accessed 2020-01-14]
36. Xie Q, Dai Z, Hovy E, Luong M, Le Q. Unsupervised data augmentation for consistency training. ArXiv 2019 epub ahead of print - 1904.12848 [FREE Full text]

Abbreviations

ADRD: Alzheimer disease and related dementias
AUC: area under the curve
EHR: electronic health record
EMR: electronic medical record
ICD: International Classification of Disease
NLP: natural language processing
OLDW: OptumLabs Data Warehouse
PPV: positive predictive value

Edited by T Hao; submitted 15.01.20; peer-reviewed by T Muto, R Gosangi; comments to author 17.02.20; revised version received 24.02.20; accepted 25.03.20; published 03.06.20.

Please cite as:
Hane CA, Nori VS, Crown WH, Sanghavi DM, Bleicher P
Predicting Onset of Dementia Using Clinical Notes and Machine Learning: Case-Control Study
JMIR Med Inform 2020;8(6):e17819
URL: https://medinform.jmir.org/2020/6/e17819
doi:10.2196/17819
PMID:32490841

Original Paper

Artificial Intelligence–Based Traditional Chinese Medicine Assistive Diagnostic System: Validation Study

Hong Zhang[1*], MSc; Wandong Ni[2*], PhD, MD; Jing Li[1*], MSc; Jiajun Zhang[3*], PhD

[1]Computer Center, Guanganmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China

[2]Certification Center of Traditional Chinese Medicine, Physician Qualification, State Administration of Traditional Chinese Medicine, Beijing, China

[3]Department of Software Engineering, NCT Lab Corp, Billerica, MA, United States

[*]all authors contributed equally

Corresponding Author:
Hong Zhang, MSc
Computer Center
Guanganmen Hospital
China Academy of Chinese Medical Sciences
5 Beixiange
Xicheng District
Beijing, 100053
China
Phone: 86 13811658952
Email: zhanghong6699@163.com

Abstract

Background: Artificial intelligence–based assistive diagnostic systems imitate the deductive reasoning process of a human physician in biomedical disease diagnosis and treatment decision making. While impressive progress in this area has been reported, most of the reported successes are applications of artificial intelligence in Western medicine. The application of artificial intelligence in traditional Chinese medicine has lagged mainly because traditional Chinese medicine practitioners need to perform syndrome differentiation as well as biomedical disease diagnosis before a treatment decision can be made. Syndrome, a concept unique to traditional Chinese medicine, is an abstraction of a variety of signs and symptoms. The fact that the relationship between diseases and syndromes is not one-to-one but rather many-to-many makes it very challenging for a machine to perform syndrome predictions. So far, only a handful of artificial intelligence–based assistive traditional Chinese medicine diagnostic models have been reported, and they are limited in application to a single disease-type.

Objective: The objective was to develop an artificial intelligence–based assistive diagnostic system capable of diagnosing multiple types of diseases that are common in traditional Chinese medicine, given a patient's electronic health record notes. The system was designed to simultaneously diagnose the disease and produce a list of corresponding syndromes.

Methods: Unstructured freestyle electronic health record notes were processed by natural language processing techniques to extract clinical information such as signs and symptoms which were represented by named entities. Natural language processing used a recurrent neural network model called bidirectional long short-term memory network–conditional random forest. A convolutional neural network was then used to predict the disease-type out of 187 diseases in traditional Chinese medicine. A novel traditional Chinese medicine syndrome prediction method—an integrated learning model—was used to produce a corresponding list of probable syndromes. By following a majority-rule voting method, the integrated learning model for syndrome prediction can take advantage of four existing prediction methods (back propagation, random forest, extreme gradient boosting, and support vector classifier) while avoiding their respective weaknesses which resulted in a consistently high prediction accuracy.

Results: A data set consisting of 22,984 electronic health records from Guanganmen Hospital of the China Academy of Chinese Medical Sciences that were collected between January 1, 2017 and September 7, 2018 was used. The data set contained a total of 187 diseases that are commonly diagnosed in traditional Chinese medicine. The diagnostic system was designed to be able to detect any one of the 187 disease-types. The data set was partitioned into a training set, a validation set, and a testing set in a ratio of 8:1:1. Test results suggested that the proposed system had a good diagnostic accuracy and a strong capability for generalization. The disease-type prediction accuracies of the top one, top three, and top five were 80.5%, 91.6%, and 94.2%, respectively.

Conclusions: The main contributions of the artificial intelligence–based traditional Chinese medicine assistive diagnostic system proposed in this paper are that 187 commonly known traditional Chinese medicine diseases can be diagnosed and a novel prediction method called an integrated learning model is demonstrated. This new prediction method outperformed all four existing methods in our preliminary experimental results. With further improvement of the algorithms and the availability of additional electronic health record data, it is expected that a wider range of traditional Chinese medicine disease-types could be diagnosed and that better diagnostic accuracies could be achieved.

(JMIR Med Inform 2020;8(6):e17608) doi:10.2196/17608

KEYWORDS

traditional Chinese medicine; TCM; disease diagnosis; syndrome prediction; syndrome differentiation; natural language processing; NLP; artificial intelligence; AI; assistive diagnostic system; convolutional neural network; CNN; machine learning; ML; BiLSTM-CRF

Introduction

The field of machine learning has experienced unprecedented and rapid development in recent years; this growth can be attributed to three factors—advanced artificial neural network architecture and algorithms, enhanced computing power, and the availability of vast amounts of training data. Machine learning has been successfully applied to many fields including medical health systems. Applications of machine learning in medical health systems can be roughly classified into two categories—image-based such as radio imaging analysis and text-based such as electronic health record analysis using natural language processing. Numerous reports have shown strong performance of image-based machine learning applications [1-6] while the successful development of text-based medical applications [7,8] remains a challenge because of its unstructured and diverse form of input data. In this age of digital medicine (and its associated deluge of digital information), it has become a daunting task for medical experts to fully utilize medical history and the test result data in a timely fashion; therefore, it is not only possible but necessary that machine learning is used to assist medical professionals in diagnostic and treatment decision making. Systems that can be used to assist medical professionals in this decision making are often called assistive diagnostic systems.

Assistive diagnostic systems have become an intense research focus for both medical practitioners and scientists in the past decade. A typical assistive diagnostic system consists of a functionality that extracts critical clinical information such as symptoms from electronic health record, and another functionality that performs deductive reasoning to predict or diagnose biomedical diseases based upon the extracted clinical information. Liang et al [9] reported an artificial intelligence–based pediatric disease diagnostic system that demonstrated high diagnostic accuracies in diagnosing common childhood diseases across multiple organ systems which was comparable to that of experienced physicians. This was accomplished by using a natural language processing technique to extract relevant symptom information from electronic health record notes, and by using logistic regression classifiers to predict the disease based upon the symptoms.

In comparison to treatment decision making in Western medicine, treatment decision making in traditional Chinese medicine is more challenging. In traditional Chinese medicine, physicians need to perform syndrome differentiation [10] as well as disease diagnosis before a decision concerning treatment can be made. A syndrome is a concept unique to traditional Chinese medicine and is an abstraction of a variety of symptoms and signs—it is a pathological summarization of a specific stage of a disease. A syndrome covers disease location, etiology, and the struggle between the disease’s pathogenic factors and the body’s resistance. In traditional Chinese medicine, the relationship between disease and syndrome is not one-to-one. Instead, disease to syndrome mapping may be considered many-to-many; therefore, the application of machine learning to decision-making processes in traditional Chinese medicine is challenging. Numerous attempts have been made to apply machine learning to traditional Chinese medicine to assist physicians in their treatment decisions [10-13]. Zhou et al [12] proposed a traditional Chinese medicine diagnostic model with multilabel classification. The model takes symptoms as input and predicts medicine disease-type and corresponding syndromes and was able to show good diagnostic accuracy. Liu et al [13] used a deep learning technique and one-versus-the-rest strategy for multilabel classification in diagnostic modeling for syndrome differentiation of traditional Chinese medicine chronic gastritis diseases and also achieved good results.

Despite these encouraging preliminary results, existing artificial intelligence–based traditional Chinese medicine systems have been limited in what their diagnostic model can diagnose (typically only one type of traditional Chinese medicine disease). In practice, it is highly desirable for an assistive diagnostic system to be capable of diagnosing or differentiating between multiple diseases and syndromes.

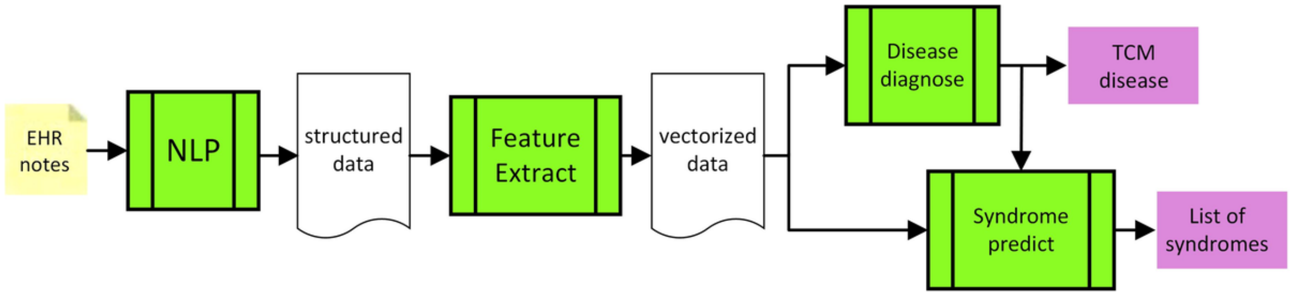
In this paper, we present an artificial intelligence–based traditional Chinese medicine diagnostic system which can diagnose 187 diseases common in traditional Chinese medicine and predict their associated syndromes from unstructured freestyle electronic health records. In the system, notes from freestyle electronic health record are first processed using a bidirectional long short-term memory network with conditional random forest [14-17] to form structured data, then features are extracted from the structured data and further vectorized. A convolutional neural network for processing text [18,19] was used to predict which traditional Chinese medicine disease was diagnosed from the vectorized data, and an integrated learning model was used to predict the disease’s corresponding syndromes.

Methods

Overview

A high-level block diagram of the diagnostic system is shown in Figure 1. The system consists of four subsystems—natural language processing, feature extraction, disease diagnosis, and the syndrome prediction. The natural language processing subsystem takes notes from freestyle electronic health record as input, extracts named entities, and produces structured data from the recognized named entities and the relationships among the named entities. The feature extraction subsystem extracts clinical information useful in disease diagnosis and syndrome differentiation from the structured data and produces additional vectorized data as output. The vectorized data are fed into a disease diagnosis network to predict the disease and are then given to the syndrome prediction subsystem to produce a list of syndromes. The syndrome prediction subsystem consists of 187 models, each of which corresponds to a disease in traditional Chinese medicine. The output of the disease diagnosis subsystem is used as input for the syndrome prediction subsystem in order for the syndrome prediction subsystem to select the appropriate model to use.

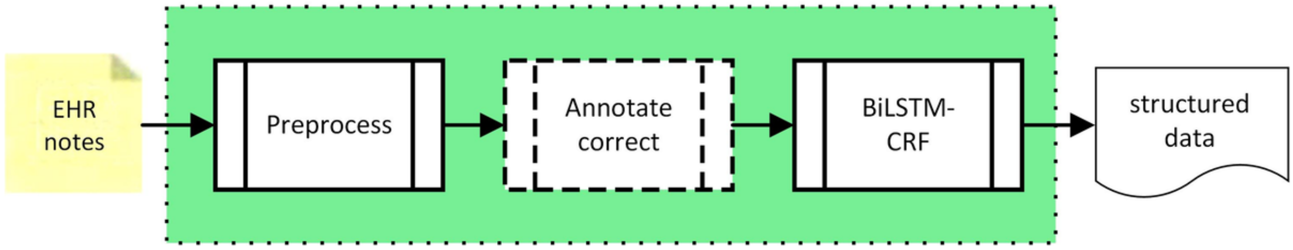
Figure 1. Block diagram of the proposed assistive diagnostic system. EHR: electronic health record; NLP: natural language processing; TCM: traditional Chinese medicine.



Natural Language Processing Subsystem

The natural language processing subsystem is responsible for generating structured data from unstructured electronic health record notes. Its internal block diagram is shown in Figure 2. There are three functional blocks in this subsystem. The first block preprocesses electronic health record notes, the second block annotates and corrects, and the third, which is a bidirectional long short-term memory network with conditional random forest, is responsible for named entity recognition. The second block exists only during the training phase of the system; during the testing and application phase, notes do not need to be annotated, thus the second block is bypassed.

Figure 2. Natural language processing subsystem block diagram. Dashed-lines indicate that the component's existence is conditional based upon whether the system is in the training phase. BiLSTM-CRF: bidirectional long short-term memory network with conditional random forest; EHR: electronic health record; NLP: natural language processing.



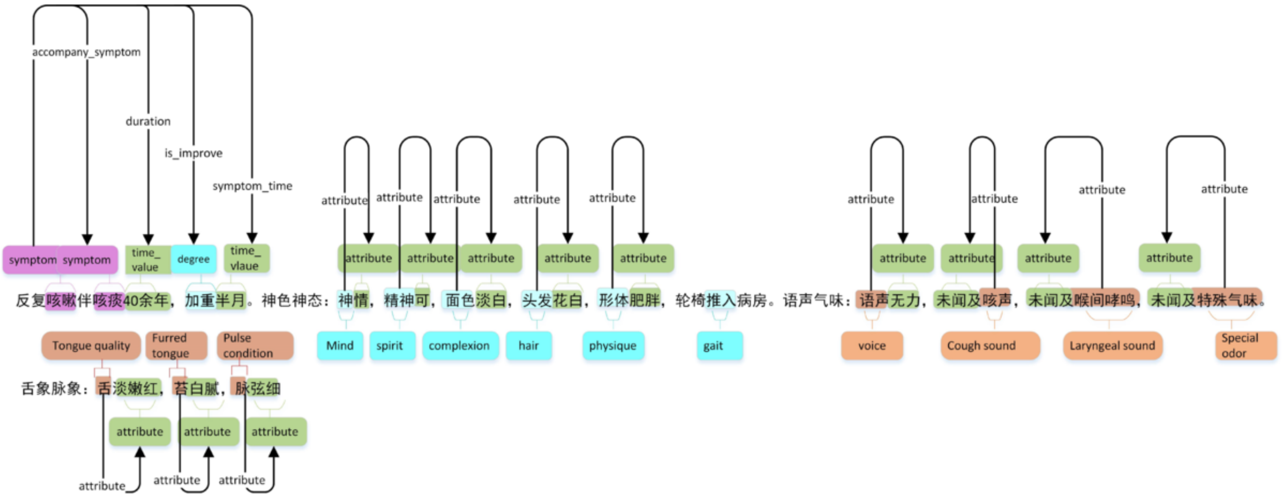
Electronic Health Record Notes Preprocessing

Electronic health record notes were preprocessed by removing unnecessary or unusable components of the electronic health record such as pictures. Notes were transformed into a standard format (half-angle encoding was used); notes were written in Chinese, and since Chinese characters can be encoded in either full-angle or half-angle format, a standard format was required. Freestyle notes were sorted and divided according to predefined sections such as chief complaint, family medical history, etc.

Electronic Health Record Notes Annotation

In the training data set, all electronic health record notes were annotated to be used for supervised training of the bidirectional long short-term memory network with conditional random forest, the convolutional neural network for processing text, and the integrated learning model network. Notes were annotated with named entities and the relationships among entities. Figure 3 shows sample annotation of the electronic patient record of a patient with a coughing history of 40 years who experienced severe coughing in the 15 days prior to visiting the physician. The electronic health record indicates that the patient entered the hospital in a wheelchair and was observed as being pale, weak, and in good spirit. Observations based upon physical examination of the tongue were recorded in the notes; tongue quality was observed as being pale red, furred, with white and greasy coating. Through annotation, the notes were marked with named entities such as cough, tongue quality, and pulse observation. Clinical information contained in the electronic health record notes was first processed by computer to form the initial training data. Subsequently, medical experts manually examined and corrected the preliminary results to form the final training data set.

Figure 3. Example of electronic health record notes annotation.



Bidirectional Long Short-Term Memory Network With Conditional Random Forest Network for Named Entity Recognition

The bidirectional long short-term memory network with conditional random forest was responsible for generating structured data from the preprocessed electronic health record notes. This was accomplished by employing a recurrent neural network as shown in Figure 4. Numerous studies have shown that a bidirectional long short-term memory network with conditional random forest is best suited for processing sequential data such as speech and text [14-16].The open-source implementation [17] of the model presented by Lample et al [16] was adopted for the construction of our bidirectional long short-term memory network with conditional random forest system.

With this network, named entities in the electronic health record notes can be extracted and properly placed in predefined data structures according to the relationships among the named entities. Figure 5 shows an example of the mapping between the electronic health record notes and the structured text with predefined sections.

Figure 4. Bidirectional long short-term memory network with conditional random forest block diagram. CRF: conditional random forest; LSTM: long short-term memory; NN: noun, singular speech tag; PRP: personal pronoun speech tag; VBO: verb speech tag; VBP: verb, singular present speech tag.

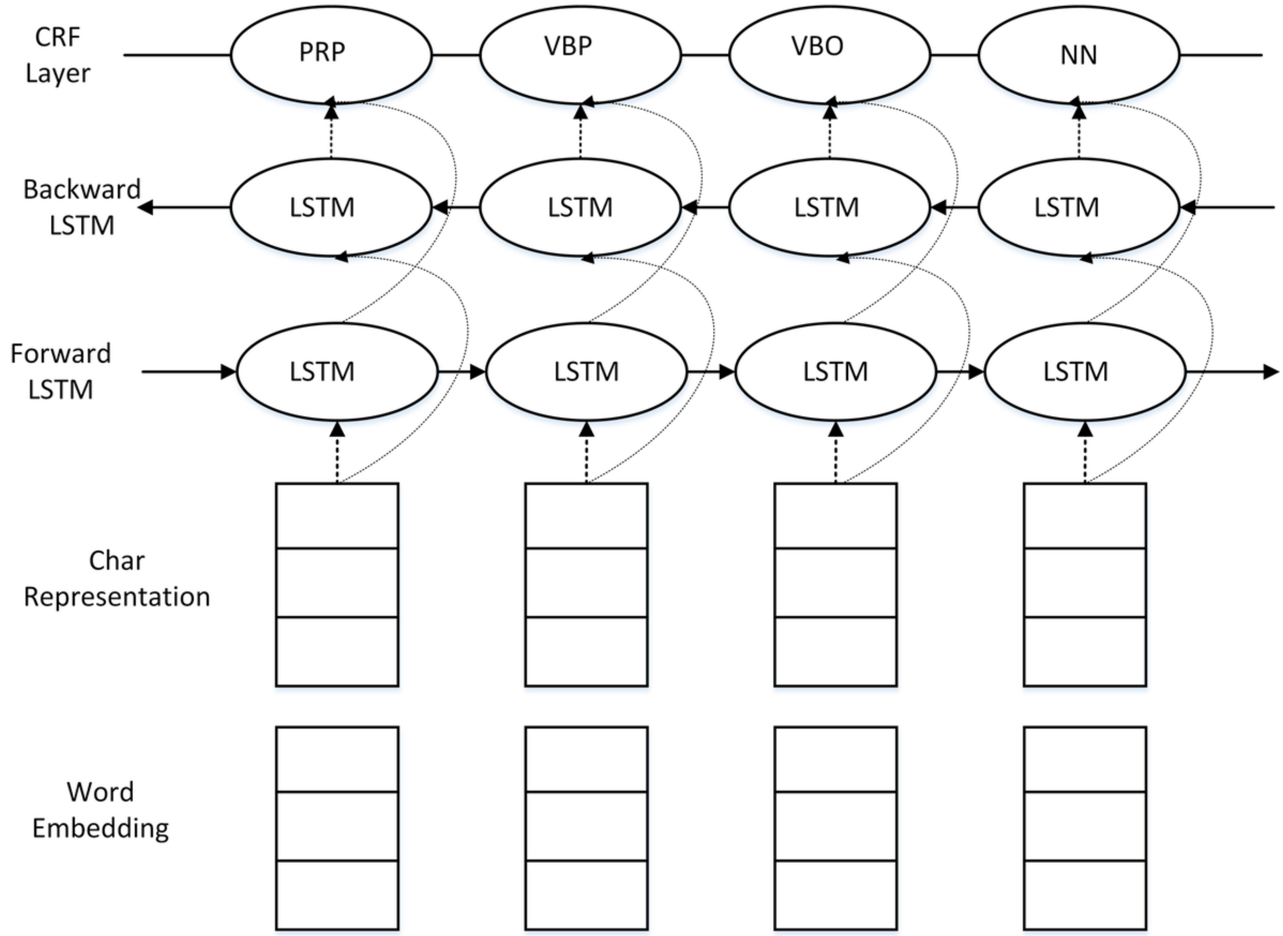
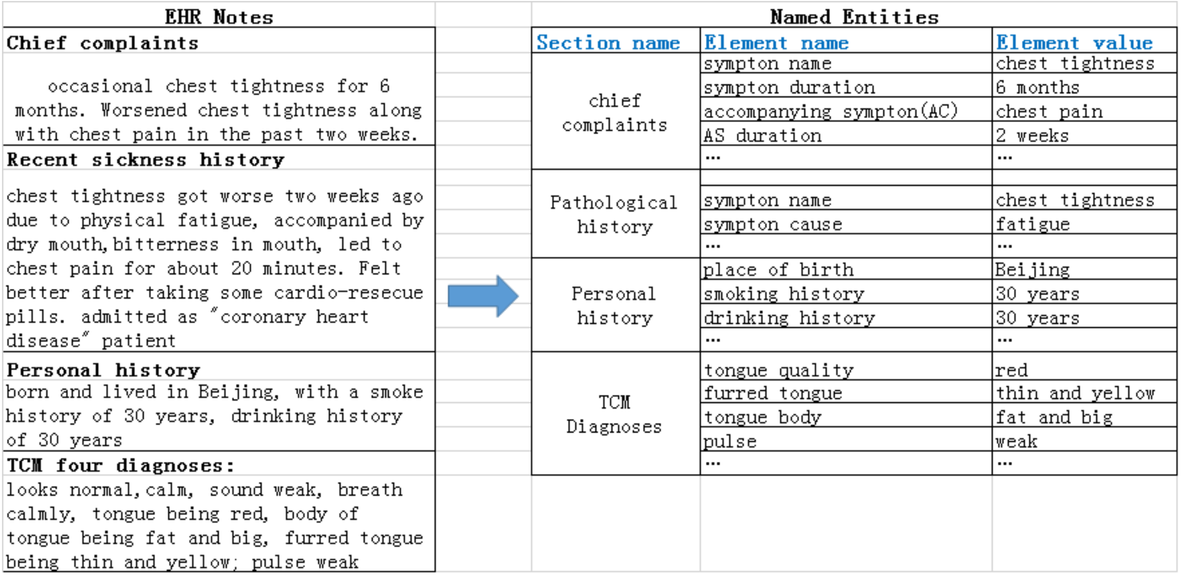


Figure 5. An example of named entity extraction from electronic health record notes. EHR: electronic health record; TCM: traditional Chinese medicine.



EHR Notes
Chief complaints
occasional chest tightness for 6 months. Worsened chest tightness along with chest pain in the past two weeks.
Recent sickness history
chest tightness got worse two weeks ago due to physical fatigue, accompanied by dry mouth, bitterness in mouth, led to chest pain for about 20 minutes. Felt better after taking some cardio-resecue pills. admitted as ″coronary heart disease″ patient
Personal history
born and lived in Beijing, with a smoke history of 30 years, drinking history of 30 years
TCM four diagnoses:
looks normal, calm, sound weak, breath calmly, tongue being red, body of tongue being fat and big, furred tongue being thin and yellow; pulse weak

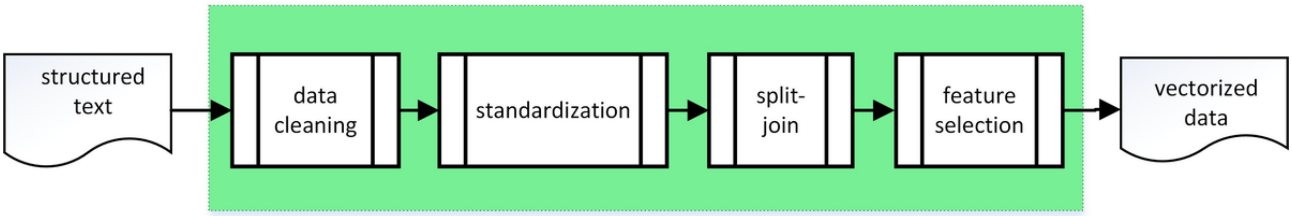
Named Entities		
Section name	Element name	Element value
chief complaints	sympton name	chest tightness
	sympton duration	6 months
	accompanying sympton(AC)	chest pain
	AS duration	2 weeks
	…	…
Pathological history	sympton name	chest tightness
	sympton cause	fatigue
	…	…
Personal history	place of birth	Beijing
	smoking history	30 years
	drinking history	30 years
	…	…
TCM Diagnoses	tongue quality	red
	furred tongue	thin and yellow
	tongue body	fat and big
	pulse	weak
	…	…

Feature Extraction Subsystem

The feature extraction subsystem was responsible for extracting useful information from the structured text data. The internal blocks of this subsystem are shown in Figure 6. The structured text contained many redundant and nonrelevant entries since the data structure was defined for a generalized purpose. During the feature extraction process data were cleaned, descriptions of symptoms and physical conditions were standardized (given that different physicians may have used different wording) using a predefined dictionary, a process called split-and-join was performed. Since the same named entity could be found in different sections of the structured text, for example, in the medical history as well as in the chief complaint, this step split the sections into parts and joined the relevant parts based upon their features.

During feature selection, entities corresponding to the same symptom were correspondingly ordered. For example, a chief complaint of "coughing for 3 days accompanied coughing phlegm for 2 days" contained two symptom entities—"coughing" and "coughing phlegm"—and two time entities—"3 days" and "2 days." In this example, a total of four symptoms were obtained— "coughing", "coughing for 3 days", "coughing phlegm," and "coughing phlegm for 2 days"— from which a weighted sum was calculated. An entity's weight was calculated based upon the time distance from the current time as $weight(n+1) = weight(n) + increment$, where n=0, 1, 2, …N, $increment$= 1/N, $weight(0)$=0 and N was the total number of time units. This formula gives a larger weight value to an entity that is nearer in time to the current time, and a smaller weight value to an entity that is further in the past. The weighted sum was used to decide which features were extracted and the extracted features were output as vectorized data.

Figure 6. Block diagram of the feature extraction subsystem.

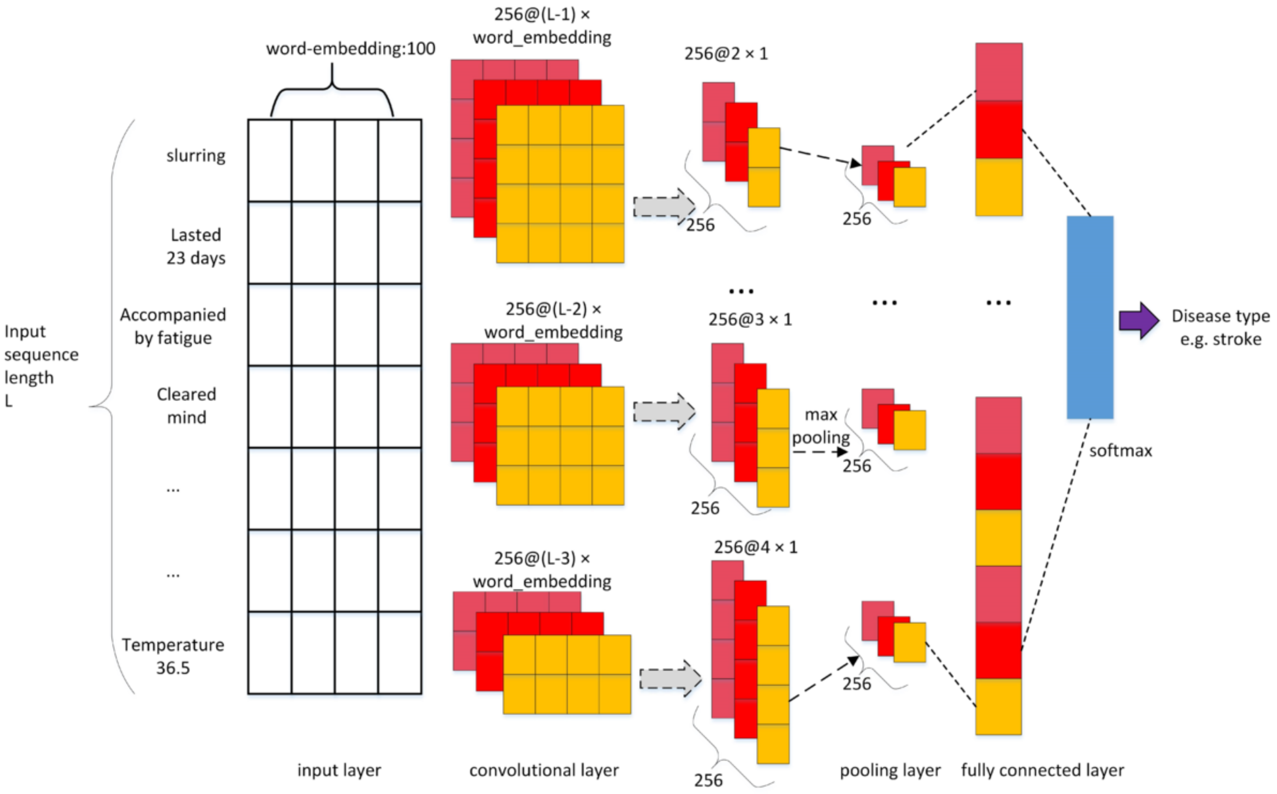


Convolutional Neural Network for Processing Text Disease Diagnosis Subsystem

Convolutional neural networks are composed of alternating convolution and pooling layers and a fully connected layer. Due to the characteristics of convolution kernel, the features represented by adjacent elements in a 2-dimensional space can be mined. Similarly, in the field of natural language processing field, 1-dimensional convolution kernels can be used to mine correlations among different words in a sentence. A network that uses convolutional neural networks for natural language text processing is called a convolutional neural network for processing text network. After word segmentation of a Chinese sequence, word embedding represented each word with a high-dimensional vector denoted by floating-point numbers to convert a sentence into a 2-dimensional matrix. Convolution operations are performed on the 2-dimensional matrix with multiple convolution kernels whose widths were equal to the dimension of the word vector dimension but which were of different heights. Pooling operations were then performed to classify and to predict Chinese text [19].

The structure of the convolutional neural network model used in this study is shown in Figure 7. The inputs were the named entities and their relationships that were extracted from the database of the structured medical record information. To ensure that the input length was consistent, the maximum number of words in the sample was set to L and zero-padding was used. From the word embedding layer, a word matrix with a size of 149,076×100 was obtained. The word vector model used in this experiment was trained by the public open-source Gensim module, whose corpus is composed of Chinese electronic health record data from multiple hospitals. The dictionary contained 149,076 words, each word represented by a 100-dimensional word vector. A 2-dimensional convolution was used in this convolutional neural network. When selecting the model structure and parameters, various factors such as sample size, hardware equipment performance, model complexity, characteristics of the electronic health records, and past experimental experience were considered. The grid search method was used to set multiple values in descending order, and to select the best parameter value from different ranges and magnitudes of the same parameter. By comparing the accuracy of the trained model with that of the test set, we set 256 convolution kernels with dimensions of (L-1)*100, (L-2)*100, and (L-3)*100 (as filters), from which 256 2×1, 3×1, and 4×1 feature surfaces were obtained. A maximum pooling layer was added to perform dimension reduction on the features of the filter layer. Finally, the pooled vectors were stitched through the fully connection layer as the input of the softmax (normalized exponential) layer to predict the disease from 187 possible classes. Due to the complexity of the multiclass classification problem and because the input electronic health record data may not be ideal, the prediction accuracy of the model in the first class cannot reach accuracy of 100%. The top five classes can be predicted accurately and also have practical significance; therefore, the top five are used in the model prediction as the final output.

Figure 7. Illustration of the text convolutional neural network.



Integrated Model for Syndrome Prediction Subsystem

Syndrome differentiation is an integral part of treatment in traditional Chinese medicine. The syndrome prediction subsystem produces a list of the most probable syndromes based upon the structured vector data. In theory, many machine learning algorithms could be used for syndrome prediction; however, in practice, not all are suited to the task due to the characteristics of the relationships between disease and its associated syndromes in traditional Chinese medicine. For example, text processing convolutional networks were ruled out since they cannot perform well in a situation where the number of syndromes associated with a disease is small. Back propagation [20] neural networks have strong nonlinear mapping capabilities because they can approach arbitrarily close to any continuous curve. Furthermore, back propagation possesses flexibility in terms of the number of network layers, the number of neurons in a layer, and the learning rate coefficients. Thus, back propagation networks have been favored in traditional Chinese medicine modeling [20]. The support vector classifier [21] has a strong mathematical basis and has shown excellent performance in situations where the number of samples is small, the dimension is high, and there is strong nonlinearity which is why support vector classifiers have previously been used for syndrome prediction. Random forest [22] models have also been used for syndrome prediction. Random forests employ bootstrap aggregation ensemble methods which combine the predictions from multiple independent decision trees. Extreme gradient boosting [23] has recently become popular and has proven to be effective in syndrome prediction.

A closer examination of these four algorithms for syndrome predictions demonstrated that, individually, they are prone to either underfitting or overfitting in applications of syndrome prediction. In our system, they were collectively employed to form an integrated learning model for a given disease-type. In this integrated learning model, the bootstrap aggregation (random forest) ensemble method was used to combine the predictions from different methods such as back propagation and support vector classifier. The extreme gradient boosting was used to combine weak classifiers into a strong classifier. Figure 8 illustrates the 187 integrated models, each of which can produce a list of syndromes for a given disease.

Each of the models consists of four individual algorithms—back propagation, support vector classifier, random forest, and extreme gradient boosting. As shown in Figure 9, the integrated model selects the final output from the outputs of the four algorithms by majority-rule. In *majority-rule*, the selection decision is based upon the highest number of votes for the outputs from each of the four algorithms. This approach not only overcomes the drawbacks of underfitting and overfitting, but also takes advantage of the strength of individual algorithms in predicting syndromes for some but not other types of diseases.

The integrated learning model approach has a better capability for generalization compared to the capabilities of existing approaches. This allows our artificial intelligence–based assistive diagnostic system to handle 187 classes of disease while existing systems may only be capable of handling one or two.

Figure 8. Illustration of the 187 integrated models.

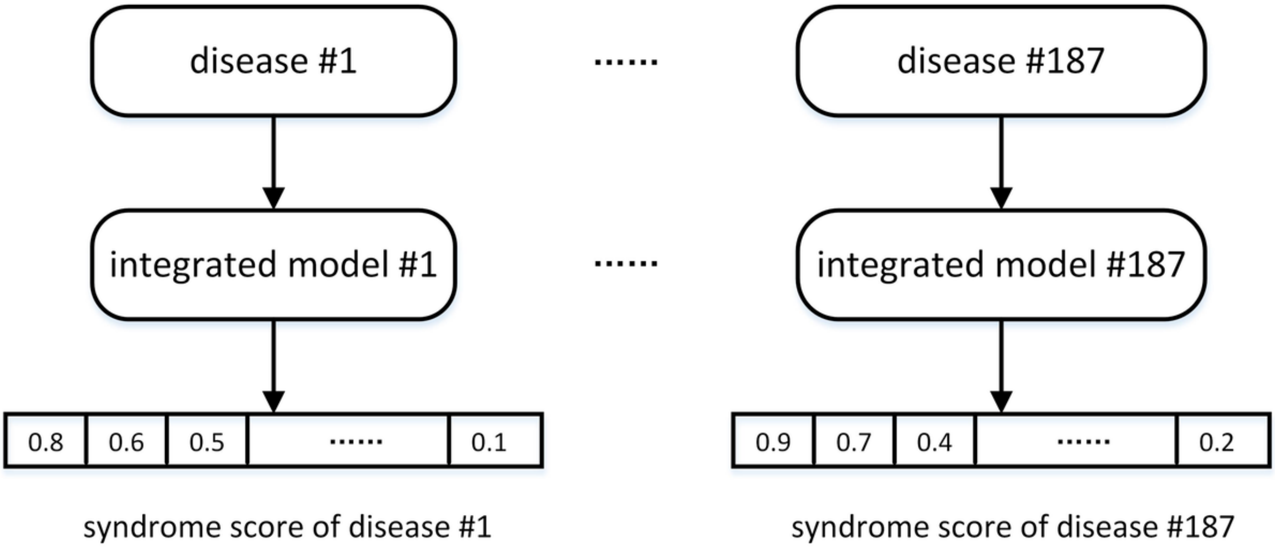
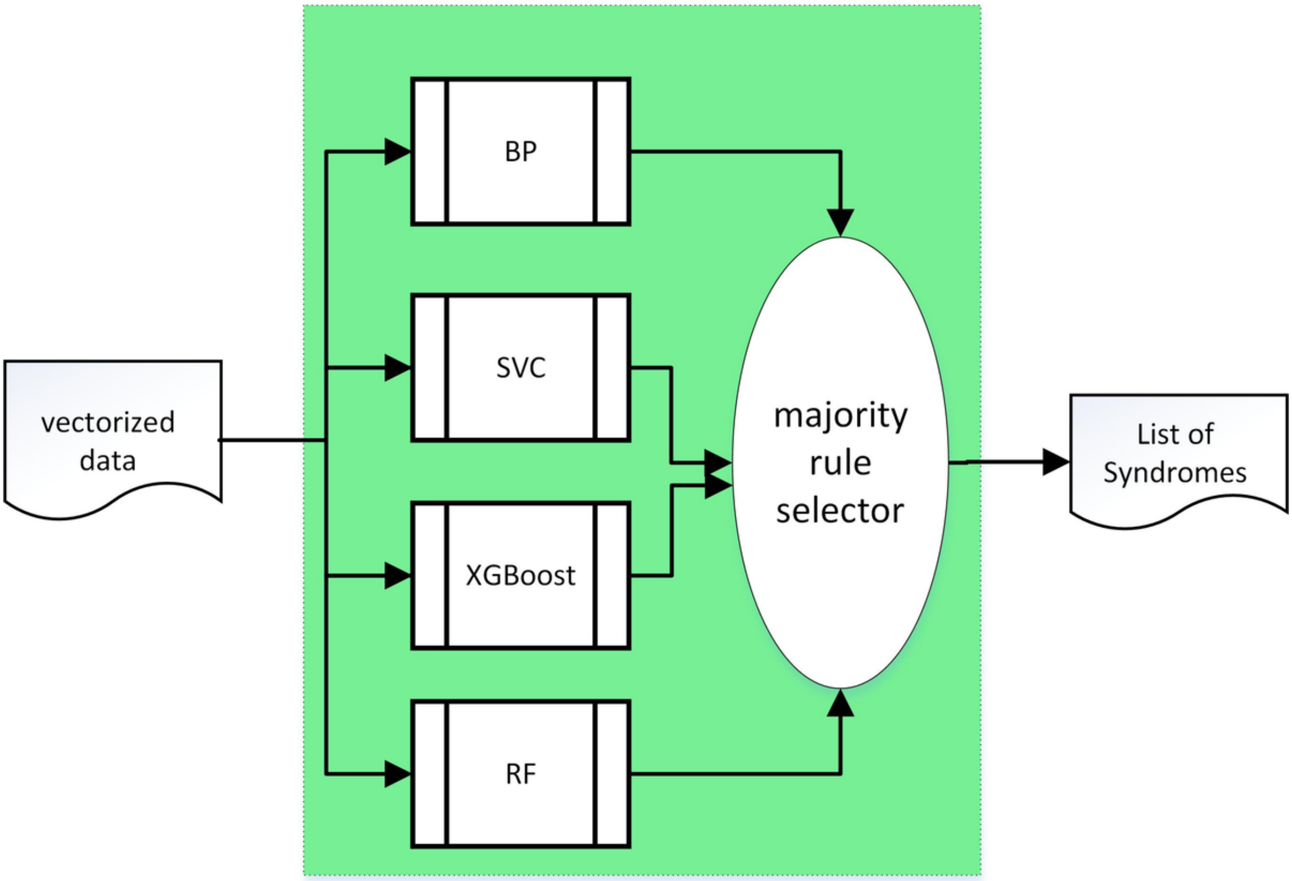


Figure 9. Block diagram of the Integrated learning model. BP: back propagation; RF: random forest; SVC: support vector classifier; XGBoost: extreme gradient boosting.



Results

Data Source

The data set used in this research was obtained from Guanganmen traditional Chinese medicine Hospital in Beijing, China. A total of 22,984 electronic health record notes that were generated between January 1, 2017 and September 7, 2019 were used for the training, validation, and testing of this system. This data set contained 187 first-category diseases and a total of 466 first-category syndromes. Furthermore, these 187 traditional Chinese medicine diseases are among the 236 most common diseases in traditional Chinese medicine [24]. These diseases cover internal medicine, gynecology, pediatric, orthopedics and traumatology, otolaryngology, dermatology, and surgery.

Originally, there were 23,719 electronic health record notes. A quality control process was applied to exclude notes with incomplete records such as missing admission page or discharge page, or notes with inconsistent information such as conflicting information between admission page and discharge page. In addition, notes that did not contain standard descriptions of complaint were discarded to eliminate biased or incorrect opinions from different physicians. From this process, 735 notes were discarded resulting in a total of 22,984 notes that were included.

The distribution of the number of electronic health record notes for different diseases and syndromes was 2180, 1913, 109, and 584 notes for cluster disease, diabetes, asthma, and spleen disease, respectively. Since this distribution imbalance would lead to bias that favors those with a large number of training samples and would reduce the system’s generalization capability, to mitigate this issue, upsampling and downsampling were used to preprocess the original data set in order to make sample distribution approximately even. Upsampling with the synthetic minority oversampling technique was used to increase the number of electronic health record notes for asthma and spleen disease to 1000 each, while the number of electronic health record notes for cluster disease and diabetes were each trimmed to 1000 through downsampling.

The processed data set was then partitioned into the training set, the validation set, and the testing set in a typical ratio of 8:1:1. The training set was used to train the coefficients of the models, the validation set was used for adjusting the model parameters, and the test set was used for measuring the performance of the system. During the partition of the data set, a k-fold+bootstrap resampling technique was employed to process the training set and the validation set. Generalization capability was improved by searching the best superparameters on different partitions during the training and integration of multiple models.

Validation

One-tenth of the total number (2298/22,984) of electronic health record in the data set was used for validation. The purpose of validation was to fine tune the neural network parameters.

Disease Diagnosis Results

The convolutional neural network for processing text disease diagnostic system was trained with a data set that contains 187 types of traditional Chinese medicine diseases. The test data set contained 2298 copies of electronic health record notes. The test results on the trained convolutional neural network for processing text model were 83.9% for the top 1 score, 92.4% for the top 3 score, and 95.7% for the top 5 score. As indicated by the test results, disease diagnosis generated relatively high diagnostic accuracies for the top 1, top 3, and top 5 score. The top 1 score, the top 3 score, and the top 5 score were calculated as follows. First, the list of predicted diseases was sorted into descending order based upon associated probabilities. If the number one predicted disease matched the target disease, then the test was considered a success for the top 1 score. If one of the first three predicted diseases matched the target disease, then the test was considered a success for the top 3 score. If one of the first five predicted diseases matched the target disease, then the test was considered a success for the top 5 score.

Based on the above definitions for the top 1 score, the top 3 score, and the top 5 score, naturally, it is always valid that the top 5 score > the top 3 score > the top 1 score. The results also suggest that even for only the top 1 score, prediction accuracy is high.

Syndrome Prediction Results

For each traditional Chinese medicine disease, a syndrome prediction model of that disease was trained. Under the same experimental conditions and data set, a 5-fold cross-validation method was used to evaluate the prediction results of each of the four algorithms and that of the integrated learning model. The calculated prediction accuracies of all 187 traditional Chinese medicine diseases were calculated. For the sake of brevity, only the results of 12 disease were included in Table 1 along with the average accuracies over all 187 diseases.

As shown in Table 1, extreme gradient boosting generally outperformed back propagation, support vector classifier, and random forest methods. Furthermore, the integration learning model reached an average prediction accuracy of 0.91, better than any of the four known models. The reason for the outstanding performance of the integration model lies in the fact that it employed a majority-rule selection method for the final prediction result.

Table 1. Syndrome prediction results.

Syndrome	Model Accuracy				
	Back Propagation	Support Vector Classifier	Random Forest	Extreme gradient boosting	Integration
Medical condition					
Bloody Stool	0.872	0.868	0.886	0.890	0.942
Abdominal Pain	0.808	0.800	0.822	0.880	0.877
Cough	0.832	0.872	0.894	0.896	0.935
Stroke	0.823	0.852	0.892	0.903	0.912
Insomnia	0.841	0.925	0.986	0.950	0.984
Hemoptysis	0.802	0.799	0.932	0.919	0.972
Blindness	0.784	0.779	0.790	0.846	0.803
Depression	0.809	0.817	0.819	0.820	0.826
Asthma	0.953	0.879	0.951	0.950	0.954
Anorectal Disease	0.908	0.861	0.879	0.901	0.893
Pepey Disease	0.832	0.865	0.870	0.887	0.892
Mean accuracy (of all 187)	0.822	0.872	0.868	0.886	0.913

Assistive Traditional Chinese Medicine Diagnostic System

Existing assistive traditional Chinese medicine diagnostic systems that have been reported can handle only one type of disease or a few syndrome predictions. Our system can diagnose 187 traditional Chinese medicine diseases and the associated syndromes. So far, we have not found systems similar to ours.

The overall system-level prediction accuracy was calculated by dividing the total number of correct predictions by the number of test cases. A correct prediction was defined by both disease and syndrome having been correctly predicted, simultaneously.

The test results of system-level accuracy were 80.5% for the top 1 score, 91.6% for the top 3 score, and 94.2% for the top 5 score. Our disease diagnosis model and syndrome prediction model together yielded relatively high diagnostic accuracies for the top 1 score, top 3 score, and top 5 score.

Discussion

Principal Results

Unlike most previous research projects which have typically focused on traditional Chinese medicine syndrome prediction for only one disease-type, we successfully used machine learning to simulate hypothetical deductive reasoning similar to that of human physicians in order to diagnose traditional Chinese medicine disease and corresponding syndromes for 187 types of diseases. The state-of-art syndrome prediction accuracy was obtained by employing a new syndrome prediction model. The prediction accuracy of this system is sufficient to assist traditional Chinese medicine practitioners in their daily clinical work.

Limitations

The data set, which only spanned two years in a single traditional Chinese medicine hospital, was relatively small. Not all common diseases and syndromes were contained in this data set; therefore, additional clinical data are needed to further improve the system.

Comparison With Prior Work

Prior work reported by other researchers mainly focused on one particular type of traditional Chinese medicine disease [20-23]. Our work is centered around the capability of diagnosing all the traditional Chinese medicine diseases and associated syndromes. At present, the proposed system can diagnose 187 out of 236 common traditional Chinese medicine diseases.

Conclusions

Artificial intelligence in diagnosing patients is highly desirable in today's digital medical age. With an abundance of medical information contained in freestyle medical health record notes, machine learning–based assistive systems can mine the medical data to extract useful and logical information and form preliminary opinions on diseases and treatment plans. For traditional Chinese medicine, syndrome prediction is also part of the diagnostic process. Because traditional Chinese medicine diseases can be linked to many syndromes, and syndromes can be linked to many diseases, disease diagnosis and syndrome prediction are more challenging in traditional Chinese medicine than in Western medicine. An effective artificial intelligence–based traditional Chinese medicine assistive diagnostic system was developed in this research by employing bidirectional long short-term memory network with conditional random forest for named entity recognition, a convolutional network for text processing for disease diagnosis, and an integrated learning model for syndrome prediction. The main contribution of this paper is a novel syndrome prediction scheme and convolutional network that represents 187 traditional Chinese medicine diseases. The system was trained, validated, and tested by using the data set obtained from nearly 23,000 electronic health record notes from Guanganmen Hospital. The proposed system distinguishes itself from existing assistive systems in that it can predict traditional Chinese medicine disease-type and syndromes simultaneously, and it can diagnose 187 types of common traditional Chinese medicine diseases. Furthermore, preliminary results suggest that the system achieved higher prediction accuracy than all existing systems [20-23]. Future work will include optimizing the convolutional network for processing text to learn all 236 common traditional Chinese medicine diseases, further improvement of the integrated learning model for syndrome prediction, and the use of additional electronic health record notes to train the system.

Conflicts of Interest

None declared.

References

1. Jalal S, Nicolaou S, Parker W. Artificial Intelligence, Radiology, and the Way Forward. Can Assoc Radiol J 2019 Feb 01;70(1):10-12. [doi: 10.1016/j.carj.2018.09.004]
2. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell 2018 Feb;172(5):1122-1131.e9. [doi: 10.1016/j.cell.2018.02.010]
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Jan 25;542(7639):115-118. [doi: 10.1038/nature21056]
4. Chartrand G. A Primer for radiologists. Radiographics 2017;37:2113-2131 [FREE Full text]
5. Nagpal K. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. Nature 2019:2-48 [FREE Full text] [doi: 10.1038/s41746-019-0196-8]
6. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. RadioGraphics 2017 Mar;37(2):505-515. [doi: 10.1148/rg.2017160130]
7. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 2018 May 8;1(1). [doi: 10.1038/s41746-018-0029-1]
8. Miao S, Dong X, Zhang X, Jing S, Zhang X, Xu T, et al. Detecting pioglitazone use and risk of cardiovascular events using electronic health record data in a large cohort of Chinese patients with type 2 diabetes. Journal of Diabetes 2019 Feb 05;11(8):684-689. [doi: 10.1111/1753-0407.12894]
9. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 2019 Mar;25(3):433-438. [doi: 10.1038/s41591-018-0335-9] [Medline: 30742121]
10. Jiang M, Lu C, Zhang C, Yang J, Tan Y, Lu A, et al. Syndrome differentiation in modern research of traditional Chinese medicine. J Ethnopharmacol 2012 Apr 10;140(3):634-642 [FREE Full text] [doi: 10.1016/j.jep.2012.01.033] [Medline: 22322251]
11. Chen L. A survey on TCM diagnosis models. Lishizhen medicine and materia research (in Chinese) 2016:688-690.
12. Zhou L. Traditional Chinese medicine diagnosis model building based on multi-label classification. 2nd International Conference on Electronic Information Technology and Computer Engineering 2018;232:02026 [FREE Full text] [doi: 10.1051/matecconf/201823202026]
13. Liu G, Yan J, Wang Y, Zheng W, Zhong T, Lu X, et al. Deep Learning Based Syndrome Diagnosis of Chronic Gastritis. Computational and Mathematical Methods in Medicine 2014;2014:1-8. [doi: 10.1155/2014/938350]
14. Huang Z. arXiv:1508.01991. 2015 Aug. Bidirectional LSTM-CRF models for sequence tagging URL: https://arxiv.org/pdf/1508.01991v1.pdf [accessed 2020-01-02]
15. Poostchi H. BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: the First Entity-Annotated Persian Dataset. 2018 Presented at: Proceedings of the Eleventh International Conference on Language Resources and Evaluation; may 11, 2018; Miyazaki, Japan p. 2018.
16. Lample G. Neural architectures for named entity recognition. 2016 Jun 12 Presented at: The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HCT-NAACL), San Diego, California, USA.?270; june 12, 2016; san diego, california p. 260. [doi: 10.18653/v1/n16-1030]
17. github. URL: https://github.com/liu-nlper/NER-LSTM-CRF [accessed 2019-12-25]
18. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. IEEE 1998 Nov 11;86(11):2278-2324. [doi: 10.1109/5.726791]
19. Xu F. Investigation on the Chinese text sentiment analysis based on convolutional neural networks in deep learning. Comput. Mater. Contin 2019;58(3):697-709. [doi: 10.32604/cmc.2019.05375]
20. Wu L. Research on a conjugate gradient descent algorithm based BP neural network for TCM diabetic disease diagnosis model. Computer knowledge and technology (in Chinese) 2019;5(23):218-221.
21. Xu M. Application of support vector machine in the diagnosis of hypertension in TCM syndrome. Chinese Medicine (in Chinese) 2017 Jun:171-174.
22. Sun G. Cervical Cancer Diagnosis based on Random Forest. IJPE 2017. [doi: 10.23940/ijpe.17.04.p12.446457]

23. Chen R. A study on diagnosing prediction model of hypertension identified as overabundant liver-fire syndrome based on XGboost algorithm. In: Thesis, Chinese Academy of Chinese Medicine (in Chinese). Beijing, China: Chinese Academy of Chinese Medicine; 2018.
24. Symptoms in Chinese medicine: Guidelines for diagnosis and treatment of common internal diseases in Chinese medicine. Beijing: China Traditional Chinese Medicine Press; 2008.

Abbreviations

BiLSTM-CRF: bidirectional long short-term memory network with conditional random forest
CRF: conditional random forest
EHR: electronic health record
LSTM: long short-term memory
NN: noun, singular speech tag
PRP: personal pronoun speech tag
VBO: verb speech tag
VBP: verb, singular present speech tag
NLP: natural language processing

Edited by T Hao, B Tang, Z Huang; submitted 25.12.19; peer-reviewed by J Jiang, J Xu; comments to author 14.02.20; revised version received 10.04.20; accepted 19.04.20; published 15.06.20.

Please cite as:
Zhang H, Ni W, Li J, Zhang J
Artificial Intelligence–Based Traditional Chinese Medicine Assistive Diagnostic System: Validation Study
JMIR Med Inform 2020;8(6):e17608
URL: http://medinform.jmir.org/2020/6/e17608/
doi:10.2196/17608
PMID:32538797

Original Paper

End-to-End Models to Imitate Traditional Chinese Medicine Syndrome Differentiation in Lung Cancer Diagnosis: Model Development and Validation

Ziqing Liu[1,2*]; Haiyang He[3*], MA; Shixing Yan[3], MA; Yong Wang[4], BA; Tao Yang[2], PhD, MD; Guo-Zheng Li[1], PhD

[1]Second School of Clinic Medicine, Guangzhou University of Chinese Medicine, Guangzhou, China

[2]School of Artifical Intelligence and Information Techology, Nanjing University of Chinese Medicine, Nanjing, China

[3]Shanghai Bright AI Co, Ltd, Shanghai, China

[4]Shanghai Literature Institute of Traditional Chinese Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Tao Yang, PhD, MD
School of Artifical Intelligence and Information Techology
Nanjing University of Chinese Medicine
Nanjing
China
Phone: 86 13405803341
Email: taoyang1111@126.com

Abstract

Background: Traditional Chinese medicine (TCM) has been shown to be an efficient mode to manage advanced lung cancer, and accurate syndrome differentiation is crucial to treatment. Documented evidence of TCM treatment cases and the progress of artificial intelligence technology are enabling the development of intelligent TCM syndrome differentiation models. This is expected to expand the benefits of TCM to lung cancer patients.

Objective: The objective of this work was to establish end-to-end TCM diagnostic models to imitate lung cancer syndrome differentiation. The proposed models used unstructured medical records as inputs to capitalize on data collected for practical TCM treatment cases by lung cancer experts. The resulting models were expected to be more efficient than approaches that leverage structured TCM datasets.

Methods: We approached lung cancer TCM syndrome differentiation as a multilabel text classification problem. First, entity representation was conducted with Bidirectional Encoder Representations from Transformers and conditional random fields models. Then, five deep learning–based text classification models were applied to the construction of a medical record multilabel classifier, during which two data augmentation strategies were adopted to address overfitting issues. Finally, a fusion model approach was used to elevate the performance of the models.

Results: The F1 score of the recurrent convolutional neural network (RCNN) model with augmentation was 0.8650, a 2.41% improvement over the unaugmented model. The Hamming loss for RCNN with augmentation was 0.0987, which is 1.8% lower than that of the same model without augmentation. Among the models, the text-hierarchical attention network (Text-HAN) model achieved the highest F1 scores of 0.8676 and 0.8751. The mean average precision for the word encoding–based RCNN was 10% higher than that of the character encoding–based representation. A fusion model of the text-convolutional neural network, text-recurrent neural network, and Text-HAN models achieved an F1 score of 0.8884, which showed the best performance among the models.

Conclusions: Medical records could be used more productively by constructing end-to-end models to facilitate TCM diagnosis. With the aid of entity-level representation, data augmentation, and model fusion, deep learning–based multilabel classification approaches can better imitate TCM syndrome differentiation in complex cases such as advanced lung cancer.

(JMIR Med Inform 2020;8(6):e17821) doi:10.2196/17821

KEYWORDS

traditional Chinese medicine; syndrome differentiation; lung cancer; medical record; deep learning; model fusion

Introduction

Lung cancer is a source of hardship worldwide, with high incidence and mortality [1,2]. According to cancer registration data collected by the Chinese National Central Cancer Registry, over 650,000 people were diagnosed with lung cancer in 2011 [3]. Standard treatment options for lung cancer are surgery, radiotherapy, and chemotherapy [4]. However, patients with low health status, such as patients in advanced stages, tend to have low tolerability of regular treatments [5]. As a respected component of traditional Chinese medicine (TCM), Chinese herbal medicine possesses the advantages of availability, efficacy, and lower toxicity than chemotherapy and radiotherapy [6]. Moreover, its benefits and underlying mechanisms in cancer therapy have been elucidated by a body of research [7-10]. After long-term practice, clinical evidence has also shown that TCM for cancer therapy can stabilize tumor lesions, enhance quality of life, and prolong survival [11,12]. More than 1 billion TCM treatments are performed in China every year according to the China Public Health Statistical Yearbook [13], and this figure is expected to increase further; meanwhile, the number of high-level TCM experts is insufficient to support the vast need for TCM.

The efficacy of TCM treatment is based on syndrome differentiation, a diagnosis method in TCM that stratifies patients' conditions with their respective disease and then guides the choice of TCM intervention [14]. Master TCM syndrome differentiation is an intricate and time-consuming process. Because the aptitudes of clinicians vary, it can be difficult to maintain stable efficacy when treating a given disease. Therefore, differentiating syndromes when confronted with complex and aggressive cancers can be challenging [15].

From the perspective of informatics, the TCM syndrome differentiation procedure can be regarded as supervised classification. Statistical machine learning algorithms have been applied to establish TCM diagnosis models [16], such as naïve Bayes [17], decision tree [18], support vector machine [19], and K-nearest neighbor [20]. However, in clinical practice, patients can concurrently suffer from multiple diseases. In this case, TCM diagnoses of several syndromes can coexist. In this circumstance, multilabel classifiers are applied to address a problem in which a set of syndromes designates one sample. Utilizing inquiry diagnosis, Liu et al [21] constructed coronary heart disease syndrome differentiation models through various multilabel learning algorithms. Their experiment showed that the multilabel k-nearest neighbor algorithm outperformed other algorithms. Wang et al [22] formulated chronic fatigue syndrome differentiation as a multilabel learning task. Combining random forest, conformal prediction framework, and problem transformation methods, they established a reliable diagnostic tool with large-scale confidence levels from 80%-100%.

In accordance with the universal approximation theorem, a deep neural network with a given number of hidden layers should be able to approximate any function that exists between input and output [23]. With the proliferation of neural networks and the growing body of TCM clinical records, syndrome differentiation modeling approaches adopting deep neural networks have become a trend. Liu et al [24] collected 919 TCM inquiry diagnosis scales and established a deep belief network based on a multilabel model for chronic gastritis TCM syndrome diagnosis. This network demonstrated superior performance for all five evaluation measures. Moreover, the average precision was 2% higher than that of the second best performing algorithm. Xu et al [25] designed an artificial neural network with 10 hidden layers for chronic obstructive pulmonary disease TCM syndrome differentiation. According to the Global Initiative for Chronic Obstructive Lung Disease, 18,471 structured TCM outpatient medical records were separated into 4 subgroup datasets, and the subgroup artificial neural network models were trained. The evaluation indicated that subgroup syndrome differentiation models outperformed the full-group model.

Due to the flexibility and compactness of TCM clinical records, datasets used in syndrome classifier training tend to be constructed manually from free-text medical records to reproduce the syndrome differentiation process. This is a labor-intensive task that requires extensive medical expertise; some information loss is inevitable [26,27]. Considering the inaccessibility of TCM literature, Hu et al [28] modeled yin-yang syndrome differentiation as a text classification task. By employing a convolutional neural network (CNN) and the fastText classifier, two sets of experiments were conducted. The results showed that the CNN system using 5-gram characters as its inputs was the most accurate.

The aforementioned studies denote that weighted mathematical logic operation–based models can be used for intelligent TCM syndrome differentiation. However, symptom classification and the determination of diagnostic thresholds are subjective; thus, many adjustments are needed. Moreover, disputes persist regarding the objectification and correction of the weighted coefficient. Furthermore, most TCM syndrome differentiation models assume that input variables such as symptoms are mutually independent. This assumption does not conform to clinical observations.

To better generalize the experience of TCM experts, we modeled syndrome differentiation for lung cancer in the form of medical record text classification. As in previous research that seeks to uncover relationships between symptoms and herbs and between syndromes and prescriptions [29], this work models TCM syndrome differentiation for lung cancer and the procedure for TCM lung cancer diagnosis. The contributions of this work are as follows:

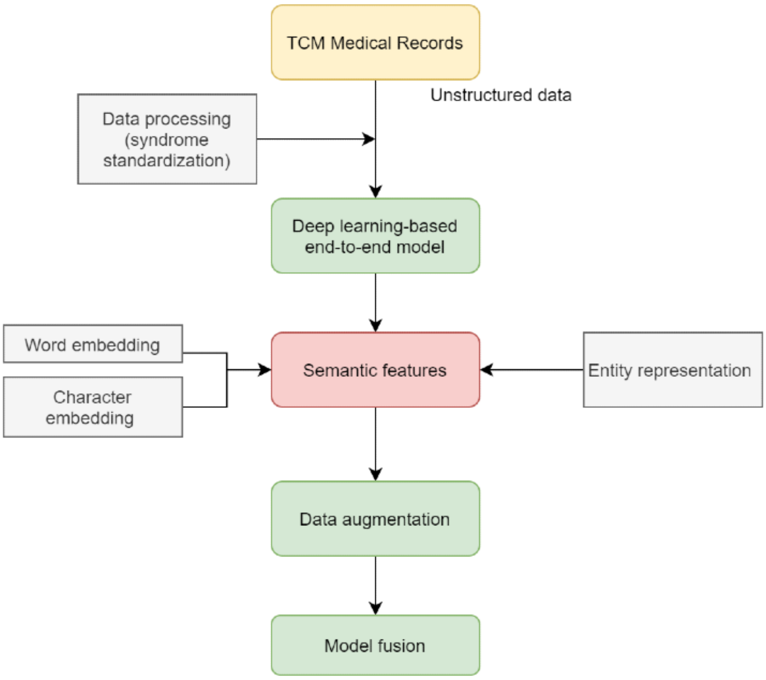
1. Syndrome factors, rather than the syndromes themselves, are adopted and standardized as labels to address the redundancy and changeability of TCM syndromes.
2. Two encoding gradients represent medical entities by applying Bidirectional Encoder Representations from Transformers (BERT) and conditional random fields (CRF) methods.
3. A data fusion approach capitalizes on all models to improve performance by building ensemble models.
4. Two data augmentation approaches were used to overcome the difficulties of ill-posed problems of samples and overfitting.

Methods

Study Design

Our work can be divided into entity-level representation learning and multilabel classifier modeling. As classified objects, TCM syndromes were split into sets of syndrome factors according to the principle of TCM syndrome factor differentiation [30]. Medical record texts were sent to the established networks to learn words and encode characters; then, the titles were extracted. Considering the difficulties of ill-posed problems of samples and overfitting, two data augmentation approaches were added. Finally, a model fusion framework was constructed. The optimum parameters for each deep learning algorithm and the best-performing algorithm were selected separately through the validation set. The framework is shown in Figure 1.

Figure 1. Framework of the end-to-end traditional Chinese medicine syndrome differentiation model.



Entity-Level Representation

We employed the BERT-CRF framework [31,32] to build entity-level representation. We used both character and word-row texts as input for the pre-trained BERT model to obtain semantic coding. We then saved it as a code list according to the word/character sequence. Meanwhile, a CRF architecture was assembled as the output layer to predict the text sequence labels and recognize the medical entities. Based on the semantic code list and the recognized entities, we generated entity-level representation with concatenating individual code in the order of the defined code list. We believed that the entity-level strategy would exploit the prior knowledge of TCM medical information that was implicitly learned during training. Multilabel classifier modelling was used for syndrome differentiation.

As shown in Figure 2, the deep learning–based syndrome differentiation models consisted of a classification layer and a sigmoid activation function. The models were fed by preprocessed TCM medical records and produced a sequence of label scores corresponding to each category. If the confidence score was higher than the threshold (ie, 0.5), the category label was added to the final syndrome differentiation.

Let $\chi = (x_1, x_2, x_3, \ldots, x_N)$ denote the N dimension sample space of a medical record text and $Y = (y_1, y_2, y_3, \ldots, y_m)$ denote the set of lung cancer syndrome factor labels. Formally, the syndrome differentiation multilabel learning task can then be defined as follows:

The multilabel task is to learn a function f: $\chi\ 2^Y$ from a given dataset $((x_1, Y_1), (x_2, Y_2), (x_3, Y_3), \ldots, (x_N, Y_N))$, where $x_i \in \chi$ and $Y_i \subseteq Y$ are the m-dimension label sets.

The universal approximation theorem indicates that a feed-forward deep network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets under mild assumptions on the activation function [33]. In our experiment, the multilabel models with deep learning approximated the function f: $\chi\ 2^Y$ and obtained the syndrome factor prediction labels in lung cancer diagnosis. Our experiment used fastText, text-convolutional neural network (Text-CNN), text-recurrent neural network (Text-RNN), recurrent convolutional neural network (RCNN), and text-hierarchical attention network (Text-HAN) models to approximate the function f.

For a deep learning–based multilabel classifier, the network parameters in the label matching module must be learned from a training dataset. The classifier is represented as C. For N-class multilabel classification, we used binary cross-entropy loss function and added L2 regularization to all model parameters. The total function is as follows:

(1)

where y^*_i indicates the ground truth predictions of the *i*th sample from the training dataset, y_i, is the label of the task, Φ denotes all the parameters of the model, and λ_Φ is the regularization hyperparameter.

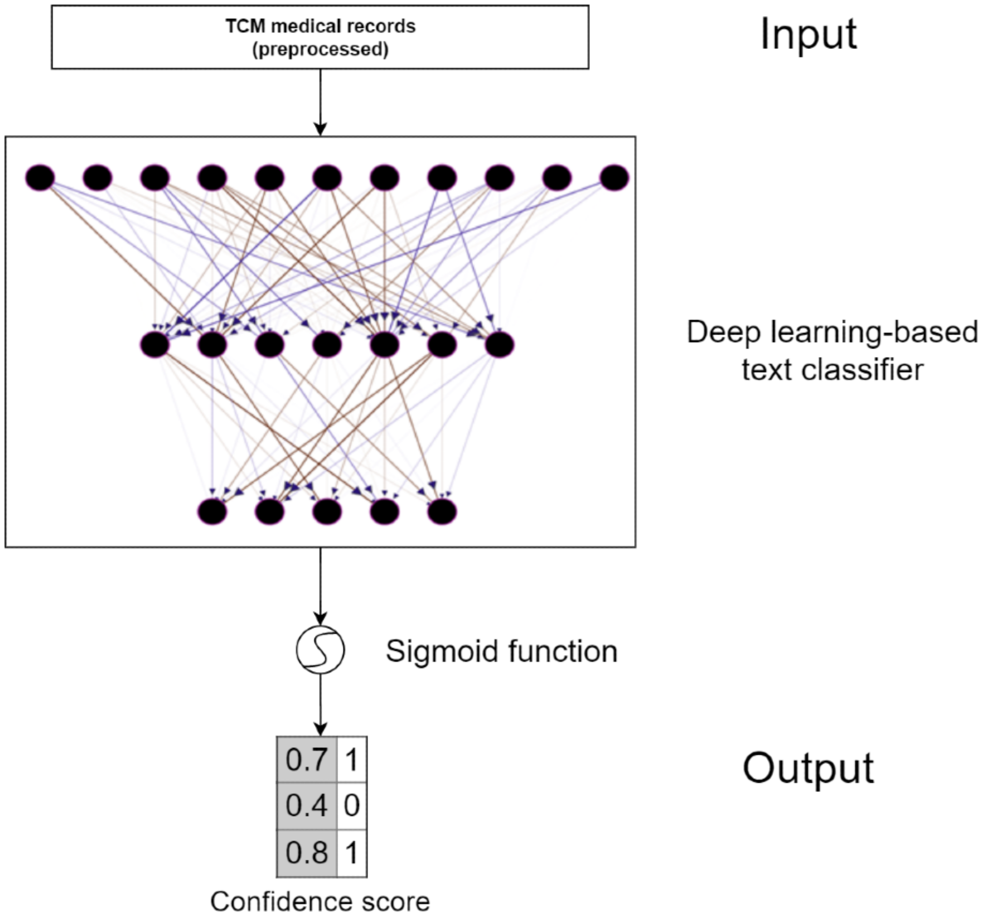
We converted the multilabel classification to multiple binary classifications. The confidence score for each label in the prediction results was then obtained with multiple logistic regression models. We employed the sigmoid activation function for each label to compute the confidence score through a linear combination of each vector as

$$score = sigmoid(w_i O)$$

(2)

where O is the output of the last layer and w_i indicates the weight. In our experiment, if the confidence score for each category was >0.5, the corresponding label was included in the prediction results. If the score was <0.5, the corresponding label was not included in the prediction results.

Figure 2. Schematic of the deep learning–based multilabel classifier.



Deep Learning–Based Classifiers

fastText [34] was used as the baseline model in our experiments. fastText is often on par with deep neural networks in terms of classification accuracy.

The first classifier was a Text-CNN model [35]. The input word was embedded to obtain a 3D sensor. Next, a convolution layer with multiple filter widths of varying sizes and pooling layers was adopted to extract local features. We then concatenated the sigmoid function with the final fully connected layer. In this way, the Text-CNN could capture partial textual features.

The text-RNN model uses bidirectional long short term memory to extract context information and global information about sentences [36]. A traditional text-RNN uses the last hidden layer as the classification. To extract context information for each word, we used k-Max pooling for all hidden elements. We then used a fully connected layer with a sigmoid function to classify the lung cancer syndromes. In this experiment, we applied a text-RNN model with N features as inputs per sentence.

In the RCNN model [37], a recurrent structure is utilized to capture as much contextual information as possible when learning word representations. This may introduce less noise than traditional window-based neural networks. We employed a convolution layer and max pooling layer to automatically judge which words were crucial in the text classification and to capture the key components in the text. Then, the lung cancer syndrome was classified using a fully connected layer with a sigmoid function.

The Hierarchical Attention Network (HAN) [38] mirrors the document’s structure. It progressively constructs a document representation by aggregating important words into sentence representation and then aggregating important sentence representation into document representation. Therefore, two bi-directional Gate Recurrent Unit (bi-GRU) models are set to acquire the varying levels of sequence encoding. Furthermore, considering the fact that the importance of words and sentences is context-dependent, two levels of attention layers are added separately after the sequence encoder. In this way, the model can vary the amount of attention to individual words and sentences when constructing the document’s representation.

Data Augmentation

To address possible overfitting, we added two data augmentation approaches (ie, we shuffled the sentence randomly and dropped words with a given probability). Consider the sentence ”胸片结果发现胸腔积液,去胸科医院排除结核” (*chest radiography examination shows pleural effusion, went to Chest Hospital to exclude the possibility of TB*). Using the shuffle method, the sentence may become “排除结核去胸科医院，结果发现胸腔积液胸片” (*to exclude the possibility of a TB patient going to the Chest Hospital, the examination shows pleural effusion chest radiography*); in the dropping method, it may become “胸片胸腔积液，胸科排除结核” (*chest radiography pleural effusion, Chest to exclude the possibility of TB*). During the model training batch, we used the shuffle mechanism and dropping mechanism to avoid overfitting and to ensure that the models demonstrated differences.

Evaluation Metrics

We used evaluation metrics to measure the performance of the learning methods in our experiment. We employed micro-averaging methods to average the classes. In this way, each class could be summed and their averages could be computed.

Precision

Precision and recall are useful prediction success evaluation metrics when a class is imbalanced. Precision is the measure of the relevancy of the results and was computed as follows:

(3)

where $f(x_i)$ is the output classifier function and y_i indicates the prediction results.

Recall

The recall is a measure of how many relevant results are returned:

(4)

where $f(x_i)$ is the output classifier function and y_i indicates the prediction results.

F1 Score

The F1 score is defined as the harmonic mean of the precision and recall:

(5)

Hamming Loss

In simplest terms, the Hamming loss is the percentage of labels that are incorrectly predicted (ie, the percentage of wrong labels). The smaller the Hamming loss value, the better the performance:

(6)

where $f(x_i)$ is the output classifier function, Δrepresents the symmetry difference between the predicted label set and the true label set, and N indicates the class number.

Mean Average Precision

The mean average precision is a score that is assigned to multilabel tasks. Its value is between 0 and 1. The higher the value, the better the performance.

(7)

Area Under the Curve

The area under the curve (AUC) is one of the most important evaluation metrics for any classification model. The AUC refers to the area under the receiver operating characteristic curve.

Results

Dataset

The dataset used in the experiment consisted of 1206 clinical records of patients diagnosed with non–small cell lung cancer. The records were collected by Professor Zhongying Zhou, a renowned TCM master with expertise in lung cancer treatment. The medical records were composed of chief complaints, anamnesis, history of present illness, lab test results, four TCM examinations, and syndrome differentiation results; each visit resulted in several TCM syndrome diagnoses. Due to redundancy, the collected syndrome set required standardization, while syndromes in the dataset had distinctive personal characteristics. This causes a mapping problem in the published TCM syndrome standards that have been prevalent for decades [39]. To preserve as much of the original diagnosis as possible, we transformed each syndrome into a set of syndrome factors. These were regarded as the assembly parts of the TCM syndromes. The feasibility of this transformation has been discussed by Luo et al [40]. The splitting followed TCM syndrome factor differentiation [30]. Before factorizing, there were nearly 600 distinctive TCM syndrome labels, with 2-4 labels for each record. When the syndromes were replaced by TCM syndrome factors, only syndrome labels were left, with 2-6 labels for each record. The 12 obtained syndrome factor labels and their frequencies are shown in Table 1.

Table 1. TCM syndrome factors for lung cancer and their frequencies.

Syndrome factor	Frequency
Yin deficiency	1069
Qi deficiency	1052
Phlegm	1036
Stasis	1035
Cancer toxin	766
Irascibility	522
Wind	294
Thirst	79
Dampness	72
Yang deficiency	27
Qi stagnation	19
Blood deficiency	6

Model Training

Our experimental results were obtained by 10-fold cross-validation. The entire dataset of 1206 medical records was randomly split into 10 subsets of equal size, each consisting of 120 medical records. In each of the 10 folds, a model was trained on 8 subsets, tested on 1 subset, and validated on the remaining subset. Then, the performance was averaged over the 10 folds.

For algorithm robustness and efficiency, we applied dropout to each pooling, highway, and long short term memory (LSTM) layer. For the base model, the dropout probability was 0.5, and the learning rate was set at 0.01-0.03. The hidden state dimensions in Bi-LSTM were 256. All fully connected layers contained 512 units. Moreover, the initialization network weights were sampled in a Gaussian distribution, and the bias was initialized to 0. The minimum batch size was set to 1024. To prevent overfitting during the training process, the L2 (0.00002) regularization was added for all model parameters, and we directly minimized the loss function using Adam stochastic optimization [41].

The above experiments were implemented using a computer equipped with 2 GeForce GTX 1080 Ti graphics processing units (Nvidia Corporation).

Experimental Process

The performance of the models without and with data augmentation is shown in Tables 2 and 3. When character encoding–based representation was used as the input, the Text-HAN, RCNN, and fastText models performed best for all indicators when data augmentation was applied. Moreover, the micro-F1 scores of all five models improved. For example, in the word-encoding RCNN results with the convergence model, the F1 of RCNN with augmentation was 0.8650%-2.41% higher than that of RCNN without augmentation. The Hamming loss of RCNN with augmentation was 0.0987%-1.8% lower than that of RCNN without augmentation. These results reveal that data augmentation methods can mitigate overfitting problems.

Comparing the models, the micro-F1 scores of the Text-HAN model reached 0.8676 and 0.8751 for the character encoding–based and word encoding–based classifications, respectively; these scores are higher than those of the other four models. This may be due to the attention mechanisms and hierarchical structure, which can overcome the diffusion problem of backpropagation gradients and can detect additional information by computing the word-level and sentence-level attention. Theoretically, Text-HAN adopts two levels of attention mechanisms and hierarchical structures; thus, it can consider additional text information and ignore less relevant content when constructing the document representation.

Observing the two representation methods, the evaluation metrics denote that the models with word-encoding representation as input performed better for all indicators except for the mean average precision without data augmentation; the mean average precision of the word encoding–based RCNN with data augmentation was 10% higher than that of the character encoding–based RCNN.

To improve the classifier performance, we applied the hybrid predicting layer by linear weight after the sigmoid layer and adopted grid search methods to obtain the best hyperparameters. The hybrid results are shown in Table 4. Compared with Table 3, the model fusion approach improved the performance, especially the F1 score of the fusion model of Text-CNN, Text-RNN, and Text-HAN. The F1 score was 0.8884, which represents the best performance among the models in the experiment. Theoretically speaking, the ensemble selection used forward stepwise selection by building optimized Text-CNN, Text-RNN, and Text-HAN ensemble models. This is because the selection of features from the ensemble learning approach can exploit the advantages of all of the models to create an optimized fusion model with superior performance.

Table 2. Character encoding–based multilabel classification results.

Model	Precision	Recall	F1 score	Hamming loss	Mean average precision	AUC[a]
Unaugmented						
fastText	0.8188	0.7923	0.8053	0.1202	0.8164	0.9211
Text-CNN[b]	0.8327	0.8342	0.8334	0.1042	0.8634	0.9472
Text-RNN[c]	0.8403	0.8240	0.8321	0.1231	0.8731	0.9021
RCNN[d]	0.8467	0.8352	0.8409	0.1005	0.8842	0.9324
Text-HAN[e]	0.8314	0.8552	0.8431	0.0990	0.8361	0.9261
Augmented						
fastText	0.8447	0.8447	0.8447	0.0990	0.8752	0.9520
Text-CNN	0.8496	0.8505	0.8500	0.1094	0.8845	0.9399
Text-RNN	0.8267	0.8650	0.8454	0.1232	0.8010	0.9321
RCNN	0.8652	0.8648	0.8650	0.0987	0.9056	0.9466
Text-HAN	0.8580	0.8774	0.8676	0.0836	0.9022	0.9602

[a] AUC: area under the curve.

[b] Text-CNN: text-convolutional neural network.

[c] Text-RNN: text-recurrent neural network.

[d] RCNN: recurrent convolutional neural network.

[e] Text-HAN: text-hierarchical attention network.

Table 3. Word encoding–based multilabel classification results.

Model	Precision	Recall	F1 score	Hamming loss	Mean average precision	AUC[a]
Unaugmented						
fastText	0.8376	0.8815	0.8590	0.040	0.8651	0.9810
Text-CNN[b]	0.8241	0.8520	0.8378	0.0990	0.8468	0.9395
Text-RNN[c]	0.8403	0.8240	0.8321	0.0960	0.8679	0.9403
RCNN[d]	0.8461	0.8659	0.8559	0.0832	0.8532	0.9321
Text-HAN[e]	0.8367	0.8505	0.8435	0.0970	0.8366	0.9260
Augmented						
fastText	0.8690	0.8760	0.8725	0.033	0.8752	0.9520
Text-CNN	0.8635	0.8338	0.8484	0.0886	0.8740	0.9479
Text-RNN	0.8377	0.8783	0.8575	0.0782	0.9052	0.9640
RCNN	0.8875	0.8548	0.8708	0.0532	0.9220	0.9632
Text-HAN	0.8648	0.8857	0.8751	0.0789	0.9210	0.9575

[a] AUC: area under the curve.

[b] Text-CNN: text-convolutional neural network.

[c] Text-RNN: text-recurrent neural network.

[d] RCNN: recurrent convolutional neural network.

[e] Text-HAN: text-hierarchical attention network.

Table 4. Fusion models for multilabel classification.

Fusion model	Precision	Recall	F1 score	Hamming loss	Mean average precision	AUC[a]
Text-CNN[b] and Text-RNN[c]	0.8898	0.8648	0.8771	0.0432	0.8836	0.9432
Text-CNN and Text-HAN[d]	0.8905	0.8732	0.8818	0.0521	0.8876	0.9524
Text-RNN and Text-HAN	0.8890	0.8635	0.8761	0.0305	0.8968	0.9687
Text-CNN, Text-RNN, and Text-HAN	0.8920	0.8890	0.8884	0.0312	0.9012	0.9618

[a] AUC: area under the curve.

[b] Text-CNN: text-convolutional neural network.

[c] Text-RNN: text-recurrent neural network.

[d] Text-HAN: text-hierarchical attention network.

Discussion

Principal Findings

Syndrome differentiation is the basis of rules, prescriptions, and medication in Chinese medicine. The results of syndrome differentiation directly influence clinical outcomes. Over the long history of medical practice in China, many syndrome differentiation methods have been proposed, such as six meridian, wei, qi, ying, and blood, three-energizer, viscera, and eight principles. These methods are interdependent and guide TCM clinical practice. However, the similarities and differences of these syndromes are difficult to distinguish, as disease conditions change constantly in clinical practice. The greater the number of methods for syndrome differentiation, the more chaotic the syndrome differentiation theory. This results in confusion regarding clinical syndrome differentiation. The establishment of a model to imitate syndrome differentiation has become an active research topic in TCM informatics. In recent years, statistics-based methods such as naïve Bayes, decision tree, and ensemble learning have been used in this field. However, these methods need to extract features from raw data in advance; this is a difficult task that directly influences the outcomes. Thus, reducing this influence and building a more reasonable model for TCM practice have emerged as new challenges in scientific research of clinical TCM.

The symptoms of advanced lung cancer patients are complex; therefore, their TCM diagnoses usually combine multiple syndromes. This combination is difficult to master. In this study, we ensembled end-to-end classification models based on deep learning to solve syndrome differentiation problems in TCM. This process did not require preexisting structured TCM medical records. In this study, we used syndrome factor sets instead of syndromes for the TCM diagnosis. This produces superior standardization of the various TCM lung cancer syndromes. On this basis, we established multilabel classifiers to accomplish lung cancer syndrome differentiation based on medical records collected by TCM expert Zhongying Zhou. During preprocessing, the entity-level strategy was explored due to its ability to capture partial textual features from context information. These features are implicitly learned during training. Finally, we integrated five deep learning models and conducted experiments to test their validity and benefit for TCM syndrome differentiation. Two data augmentation methods and model fusion strategies were utilized to address the overfitting problem.

Limitations and Future Work

There are some limitations to our research. This experiment focused on a small lung cancer dataset. Although some data reinforcement methods were used, the generated data are not authentic TCM clinical data. Thus, the ensuing effects require further validation. In the future, we plan to incorporate an attention capsule network, XLNet pretrained models, and a graph neural network for lung cancer syndrome differentiation. We also plan to popularize additional TCM syndrome differentiation datasets and applications.

Conclusion

The end-to-end models we ensembled based on deep learning can imitate syndrome differentiation from the perspective of natural language processing and may have more substantial applicability than traditional statistics-based algorithms. Therefore, these models can be embedded in TCM clinical information systems and provide clinical decision support for TCM physicians during their clinical practice, especially primary care physicians and physicians in rural areas. With the aid of our ensembled end-to-end models, TCM experiences can be learned and transferred to TCM clinical support systems, which will address the imbalance of TCM medical needs and medical supplies and provide tremendous social and economic benefit. Moreover, these end-to-end models may enable TCM institutions to efficiently transform their health record metadata into data assets.

Acknowledgments

The research in this paper was supported by the National Science Foundation of China under Grant Nos. 81674099, 81804219, and 81503499; the National Key Research and Development Program of China under Grant No. 2017YFC1703500; the Priority

Academic Program Development of Jiangsu Higher Education Institutions of Nursing under Grant No. 2019YSHL068; and the Shanghai TCM Science and Technology Innovation Project under Grant No. ZYKC201601022.

Conflicts of Interest

None declared.

References

1. Cheng TD, Cramb SM, Baade PD, Youlden DR, Nwogu C, Reid ME. The International Epidemiology of Lung Cancer: Latest Trends, Disparities, and Tumor Characteristics. J Thorac Oncol 2016 Oct;11(10):1653-1671 [FREE Full text] [doi: 10.1016/j.jtho.2016.05.021] [Medline: 27364315]
2. Zou X, Jia M, Wang X, Zhi X. Changing Epidemic of Lung Cancer & Tobacco and Situation of Tobacco Control in China. Article in Chinese. Zhongguo Fei Ai Za Zhi 2017 Aug 20;20(8):505-510 [FREE Full text] [doi: 10.3779/j.issn.1009-3419.2017.08.01] [Medline: 28855029]
3. Chen W, Zheng R, Zeng H, Zhang S, He J. Annual report on status of cancer in China, 2011. Chin J Cancer Res 2015 Feb;27(1):2-12 [FREE Full text] [doi: 10.3978/j.issn.1000-9604.2015.01.06] [Medline: 25717220]
4. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin 2019 Sep;69(5):363-385 [FREE Full text] [doi: 10.3322/caac.21565] [Medline: 31184787]
5. Islam KM, Anggondowati T, Deviany PE, Ryan JE, Fetrick A, Bagenda D, et al. Patient preferences of chemotherapy treatment options and tolerance of chemotherapy side effects in advanced stage lung cancer. BMC Cancer 2019 Aug 27;19(1):835 [FREE Full text] [doi: 10.1186/s12885-019-6054-x] [Medline: 31455252]
6. Xiang Y, Guo Z, Zhu P, Chen J, Huang Y. Traditional Chinese medicine as a cancer treatment: Modern perspectives of ancient but advanced science. Cancer Med 2019 May;8(5):1958-1975 [FREE Full text] [doi: 10.1002/cam4.2108] [Medline: 30945475]
7. Ye L, Jia Y, Ji KE, Sanders AJ, Xue K, Ji J, et al. Traditional Chinese medicine in the prevention and treatment of cancer and cancer metastasis. Oncol Lett 2015 Sep;10(3):1240-1250 [FREE Full text] [doi: 10.3892/ol.2015.3459] [Medline: 26622657]
8. Qi F, Zhao L, Zhou A, Zhang B, Li A, Wang Z, et al. The advantages of using traditional Chinese medicine as an adjunctive therapy in the whole course of cancer treatment instead of only terminal stage of cancer. Biosci Trends 2015 Feb;9(1):16-34 [FREE Full text] [doi: 10.5582/bst.2015.01019] [Medline: 25787906]
9. Wang S, Wu M, Cai C, Li M, Lu J. Autophagy modulators from traditional Chinese medicine: Mechanisms and therapeutic potentials for cancer and neurodegenerative diseases. J Ethnopharmacol 2016 Dec 24;194:861-876. [doi: 10.1016/j.jep.2016.10.069] [Medline: 27793785]
10. Liu J, Wang S, Zhang Y, Fan H, Lin H. Traditional Chinese medicine and cancer: History, present situation, and development. Thorac Cancer 2015 Sep;6(5):561-569 [FREE Full text] [doi: 10.1111/1759-7714.12270] [Medline: 26445604]
11. Liu R, He SL, Zhao YC, Zheng HG, Li CH, Bao YJ, et al. Chinese herbal decoction based on syndrome differentiation as maintenance therapy in patients with extensive-stage small-cell lung cancer: an exploratory and small prospective cohort study. Evid Based Complement Alternat Med 2015;2015:601067 [FREE Full text] [doi: 10.1155/2015/601067] [Medline: 25815038]
12. Chen S, Flower A, Ritchie A, Liu J, Molassiotis A, Yu H, et al. Oral Chinese herbal medicine (CHM) as an adjuvant treatment during chemotherapy for non-small cell lung cancer: A systematic review. Lung Cancer 2010 May;68(2):137-145. [doi: 10.1016/j.lungcan.2009.11.008] [Medline: 20015572]
13. China's Health Statistics Yearbook 2019. Peking, China: Peking Union Medical College Press; Aug 2019:197.
14. Jiang M, Lu C, Zhang C, Yang J, Tan Y, Lu A, et al. Syndrome differentiation in modern research of traditional Chinese medicine. J Ethnopharmacol 2012 Apr 10;140(3):634-642 [FREE Full text] [doi: 10.1016/j.jep.2012.01.033] [Medline: 22322251]
15. Nie J, Zhao C, Deng LI, Chen J, Yu B, Wu X, et al. Efficacy of traditional Chinese medicine in treating cancer. Biomed Rep 2016 Jan;4(1):3-14 [FREE Full text] [doi: 10.3892/br.2015.537] [Medline: 26870326]
16. Zhao C, Li G, Wang C, Niu J. Advances in Patient Classification for Traditional Chinese Medicine: A Machine Learning Perspective. Evid Based Complement Alternat Med 2015;2015:376716. [doi: 10.1155/2015/376716] [Medline: 26246834]
17. Wang Y, Yu Z, Jiang Y, Liu Y, Chen L, Liu Y. A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records. J Biomed Inform 2012 Apr;45(2):210-223 [FREE Full text] [doi: 10.1016/j.jbi.2011.10.010] [Medline: 22101128]
18. Wang Y, Ma L, Liu P. Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. Comput Methods Programs Biomed 2009 Sep;95(3):249-257. [doi: 10.1016/j.cmpb.2009.03.004] [Medline: 19380172]
19. Xia C, Deng F, Wang Y. Classification research on syndromes of TCM based on SVM. In: 2009 2nd International Conference on Biomedical Engineering and Informatics. 2009 Oct 17 Presented at: Paper presented atnd International Conference on Biomedical Engineering and Informatics; 2009; Tianjin, China. [doi: 10.1109/bmei.2009.5305418]

20. Li G, Sun S, You M, Wang Y, Liu G. Inquiry diagnosis of coronary heart disease in Chinese medicine based on symptom-syndrome interactions. Chin Med 2012 Apr 05;7(1):9 [FREE Full text] [doi: 10.1186/1749-8546-7-9] [Medline: 22475180]
21. Liu G, Li G, Wang Y, Wang Y. Modelling of inquiry diagnosis for coronary heart disease in Traditional Chinese Medicine by using multi-label learning. BMC Complement Altern Med 2010 Jul 20;10:37 [FREE Full text] [doi: 10.1186/1472-6882-10-37] [Medline: 20642856]
22. Wang H, Liu X, Lv B, Yang F, Hong Y. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional Chinese medicine. PLoS One 2014;9(6):e99565 [FREE Full text] [doi: 10.1371/journal.pone.0099565] [Medline: 24918430]
23. Cybenko G. Approximation by superpositions of a sigmoidal function. Math Control Signal 1989 Dec;2(4):303-314. [doi: 10.1007/bf02551274]
24. Liu G, Yan J, Wang Y, Zheng W, Zhong T, Lu X, et al. Deep learning based syndrome diagnosis of chronic gastritis. Comput Math Methods Med 2014;2014:938350 [FREE Full text] [doi: 10.1155/2014/938350] [Medline: 24734118]
25. Xu Q, Tang W, Teng F, Peng W, Zhang Y, Li W, et al. Intelligent Syndrome Differentiation of Traditional Chinese Medicine by ANN: A Case Study of Chronic Obstructive Pulmonary Disease. IEEE Access 2019;7:76167-76175. [doi: 10.1109/access.2019.2921318]
26. Zhou X, Peng Y, Liu B. Text mining for traditional Chinese medical knowledge discovery: a survey. J Biomed Inform 2010 Aug;43(4):650-660 [FREE Full text] [doi: 10.1016/j.jbi.2010.01.002] [Medline: 20074663]
27. Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artif Intell Med 2010;48(2-3):139-152. [doi: 10.1016/j.artmed.2009.07.012] [Medline: 20122820]
28. Hu Q, Yu T, Li J, Yu Q, Zhu L, Gu Y. End-to-End syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine. Comput Methods Programs Biomed 2019 Jun;174:9-15. [doi: 10.1016/j.cmpb.2018.10.011] [Medline: 30376987]
29. Yang K, Zhang R, He L, Li Y, Liu W, Yu C, et al. Multistage analysis method for detection of effective herb prescription from clinical data. Front Med 2018 Apr;12(2):206-217. [doi: 10.1007/s11684-017-0525-8] [Medline: 28623541]
30. Zhu W. Syndrome Differentiation via Syndrome Factors. Beijing, China: People's Medical Publishing House; 2008:197-197.
31. Souza F, Nogueira R, Lotufo R. arXiv preprint. 2019 Sep 23. Portuguese Named Entity Recognition using BERT-CRF URL: https://arxiv.org/abs/1909.10649 [accessed 2020-05-25]
32. Devlin J, Chang M, Lee K, Toutanova K. arXiv preprint. 2018 Oct 11. Bert: Pre-training of deep bidirectional transformers for language understanding URL: https://arxiv.org/abs/1810.04805 [accessed 2020-05-25]
33. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A. Ensemble selection from libraries of models. In: Association for Computing Machinery. 2004 Jul Presented at: Proceedings of the Twenty-first International Conference on Machine Learning; July 4-8, 2004; Banff, AB. [doi: 10.1145/1015330.1015432]
34. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers.: Association for Computational Linguistics; 2017 Apr 15 Presented at: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics; April 2017; Valencia URL: https://www.aclweb.org/anthology/E17-2068/ [doi: 10.18653/v1/e17-2068]
35. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).: Association for Computational Linguistics; 2014 Nov 1 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing; October 2014; Doha p. 17461-11751 URL: https://www.aclweb.org/anthology/D14-1181/ [doi: 10.3115/v1/d14-1181]
36. Liu P, Qiu X, Huang X. arXiv preprint. Recurrent neural network for text classification with multi-task learning URL: https://arxiv.org/abs/1605.05101 [accessed 2016-05-17]
37. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. In: Association for Comupting Machinery. 2015 Jan 15 Presented at: Twenty-ninth AAAI conference on artificial intelligence; 2015; Austin, TX p. 2267-2273.
38. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2016; San Diego, CA p. 1480-1489 URL: https://www.aclweb.org/anthology/N16-1174/ [doi: 10.18653/v1/N16-1174]
39. China State Bureau of Technical Supervision. China National Standard Open System. Beijing: China National Standardization Management Committee; 1997 Mar 04. China National Standard: Clinic terminology of traditional Chinese medical diagnosis and treatment--Syndromes. Webpage in Chinese URL: http://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=91C7CFD75D24C43F0BCB136C26BE6345 [accessed 2020-05-25]
40. Luo W, Wu C. A Study on Syndrome Elements of Lung Cancer. Article in Chinese. Journal of Nanjing University of Traditional Chinese Medicine 2009;25(2):95-98.

41. Kingma D, Ba J. arXiv preprint. 2014 Dec 22. Adam: A method for stochastic optimization URL: https://arxiv.org/abs/1412.6980 [accessed 2020-05-25]

Abbreviations

AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
CNN: convolutional neural network
CRF: conditional random fields
LSTM: long short term memory
RCNN: recurrent convolutional neural network
TCM: traditional Chinese medicine
Text-CNN: text-convolutional neural network
Text-HAN: text-hierarchical attention network
Text-RNN: text-recurrent neural network

Edited by T Hao; submitted 15.01.20; peer-reviewed by X Zhou, MS Aslam, T Liyuan; comments to author 25.02.20; revised version received 30.03.20; accepted 11.04.20; published 16.06.20.

Please cite as:
Liu Z, He H, Yan S, Wang Y, Yang T, Li GZ
End-to-End Models to Imitate Traditional Chinese Medicine Syndrome Differentiation in Lung Cancer Diagnosis: Model Development and Validation
JMIR Med Inform 2020;8(6):e17821
URL: https://medinform.jmir.org/2020/6/e17821
doi:10.2196/17821
PMID:32543445

Original Paper

Toward Optimal Heparin Dosing by Comparing Multiple Machine Learning Methods: Retrospective Study

Longxiang Su[1*], MD, PhD; Chun Liu[2*], PhD; Dongkai Li[1*], MD, PhD; Jie He[2*], PhD; Fanglan Zheng[2], PhD; Huizhen Jiang[3], MSc; Hao Wang[1], MD, PhD; Mengchun Gong[2], PhD, MD; Na Hong[2], PhD; Weiguo Zhu[4], MD, PhD; Yun Long[1], MD, PhD

[1]Department of Critical Care Medicine, Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical, Beijing, China

[2]Digital China Health Technologies Co Ltd, Beijing, China

[3]Department of Information Management, Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical, Beijing, China

[4]Department of General Internal Medicine/Department of Information Management, Peking Union Medical College Hospital, Peking Union Medical College & Chinese Academy of Medical, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yun Long, MD, PhD
Department of Critical Care Medicine
Peking Union Medical College Hospital
Peking Union Medical College & Chinese Academy of Medical
1 Shuaifuyuan, Dongcheng District, Beijing 100730, China
Beijing
China
Phone: 86 10 69152318
Email: ly_icu@aliyun.com

Abstract

Background: Heparin is one of the most commonly used medications in intensive care units. In clinical practice, the use of a weight-based heparin dosing nomogram is standard practice for the treatment of thrombosis. Recently, machine learning techniques have dramatically improved the ability of computers to provide clinical decision support and have allowed for the possibility of computer generated, algorithm-based heparin dosing recommendations.

Objective: The objective of this study was to predict the effects of heparin treatment using machine learning methods to optimize heparin dosing in intensive care units based on the predictions. Patient state predictions were based upon activated partial thromboplastin time in 3 different ranges: subtherapeutic, normal therapeutic, and supratherapeutic, respectively.

Methods: Retrospective data from 2 intensive care unit research databases (Multiparameter Intelligent Monitoring in Intensive Care III, MIMIC-III; e–Intensive Care Unit Collaborative Research Database, eICU) were used for the analysis. Candidate machine learning models (random forest, support vector machine, adaptive boosting, extreme gradient boosting, and shallow neural network) were compared in 3 patient groups to evaluate the classification performance for predicting the subtherapeutic, normal therapeutic, and supratherapeutic patient states. The model results were evaluated using precision, recall, F1 score, and accuracy.

Results: Data from the MIMIC-III database (n=2789 patients) and from the eICU database (n=575 patients) were used. In 3-class classification, the shallow neural network algorithm performed the best (F1 scores of 87.26%, 85.98%, and 87.55% for data set 1, 2, and 3, respectively). The shallow neural network algorithm achieved the highest F1 scores within the patient therapeutic state groups: subtherapeutic (data set 1: 79.35%; data set 2: 83.67%; data set 3: 83.33%), normal therapeutic (data set 1: 93.15%; data set 2: 87.76%; data set 3: 84.62%), and supratherapeutic (data set 1: 88.00%; data set 2: 86.54%; data set 3: 95.45%) therapeutic ranges, respectively.

Conclusions: The most appropriate model for predicting the effects of heparin treatment was found by comparing multiple machine learning models and can be used to further guide optimal heparin dosing. Using multicenter intensive care unit data, our study demonstrates the feasibility of predicting the outcomes of heparin treatment using data-driven methods, and thus, how machine learning–based models can be used to optimize and personalize heparin dosing to improve patient safety. Manual analysis and validation suggested that the model outperformed standard practice heparin treatment dosing.

(JMIR Med Inform 2020;8(6):e17648) doi:10.2196/17648

KEYWORDS

heparin; dosing; machine learning; optimization; intensive care unit

Introduction

In hospitals, intensive care units are unique in that vast amounts of information are collected and displayed by computerized systems, and that the diagnostic and treatment accuracy can profoundly affect quality of care and patient outcomes [1]. Data-driven clinical decision support systems have the potential to help clinicians optimize treatment and medication in an intensive care unit to maximize the medical effect for each individual patient [2].

Heparin is one of the most commonly used medications in intensive care units, and intravenous unfractionated heparin is a fundamental method of anticoagulant therapy. In most clinical practice guidelines, heparin dosing is based only on the patient's weight; the use of a weight-based heparin dosing nomogram is the standard practice for the treatment of thrombosis [3,4]. For patients who are obese who may not receive the appropriate heparin dose if it is determined based solely on body weight, some suggestions such as reducing the initial infusion rate [5-7] or using an adjusted body weight [8] have been reported. In clinical practice, activated partial thromboplastin time typically reflects blood coagulation level. A high activated partial thromboplastin time means that blood is clotting slowly, whereas a low activated partial thromboplastin time means that blood is clotting quickly. Typically, blood samples are drawn every 4 to 6 hours to monitor activated partial thromboplastin time, and the anticoagulation therapy outcome is measured by whether the activated partial thromboplastin time reaches the therapeutic window in a timely manner; however, the weight-based method easily leads to improper doses which demonstrate subtherapeutic or supratherapeutic activated partial thromboplastin time. In addition, the risk factors that result from inappropriate doses of unfractionated heparin are unclear. Only high initial rates of infusion, advanced age, and being female have been reported to be associated with supratherapeutic activated partial thromboplastin time [9,10]. Heparin administration guidelines regarding initial loading dose, maintenance dose and rate, and the activated partial thromboplastin time measurement intervals vary widely among institutions. Additionally, clinicians choose different heparin administration routes such as intravenous push or intravenous drip due based on the immediate circumstances and requirements of the patient.

Recently, machine learning techniques have dramatically improved the ability of computers to provide clinical decision support, resulting in the possibility of computer generated, algorithm-based heparin dosing recommendations. Multivariate logistic regression [11] and multinomial logistic regression [12] have been used to estimate heparin dosing with an accuracy of approximately 60%. Algorithms have also been used in studies [13,14] for other anticoagulants such as warfarin dose adjustments, but it was found that high intrapatient variability weakened the prediction accuracy.

For these reasons, a reliable method that can help doctors quickly predict and optimize heparin doses is urgently needed. It is necessary that modeling and prediction of the therapeutic window of activated partial thromboplastin time take into account multiple factors during patient treatment in order to provide appropriate decision support suggestions which can help guide clinicians in determining and preparing subsequent heparin doses or adjusting dose rate.

Methods

Data Set

Data were extracted from the Multiparameter Intelligent Monitoring In Intensive Care III database (MIMIC-III) [15] and e–Intensive Care Unit Collaborative Research database (eICU) [16] with the goal of comparing multiple predictive models and evaluating the results in different groups of patients. A cross-database evaluation was conducted. The MIMIC-III database and eICU database are free and open data sets containing medical data. The MIMIC-III database contains data from the intensive care unit at the Beth Israel Deaconess Medical Center and is published by the Laboratory for Computational Physiology at Massachusetts Institute of Technology. The eICU database, published by the Philips e–Intensive Care Unit Research Institute, is populated with data from a combination of many critical care units throughout the continental United States. Data were extracted from the databases for 14,806 adult patients who received heparin therapy during their stay in the intensive care unit. Only patient data with activated partial thromboplastin time measurements taken 4 to 6 hours after their initial heparin dose administration were used which reduced the cohort size to 3835. We chose 4 to 6 hours based on past experience and previous research [11]; it is the period within which the first activated partial thromboplastin time measurement typically occurred for the greatest proportion of patients. In clinical practice, there are different administration routes to deliver medication. Both intravenous push and intravenous drip are commonly used to deliver heparin, and in practice, are chosen based on patient condition and doctor preference; therefore, patient data were further classified by administration route—intravenous push (data set 1) and intravenous drip (data sets 2 and 3).

Feature Selection

The outcome of interest was activated partial thromboplastin time 4 to 6 hours after initial heparin infusion. Since the data were from the Beth Israel Deaconess Medical Center, we applied the definition of therapeutic time used at Beth Israel Deaconess Medical Center for the definition of therapeutic time of activated partial thromboplastin time in this study to ensure consistency. Normal therapeutic was defined as activated partial thromboplastin times from 60 seconds to 100 seconds, supratherapeutic was defined as activated partial thromboplastin times greater than 100 seconds, and subtherapeutic was defined as activated partial thromboplastin times less than 60 seconds [11]. Clinical features of interest were selected to optimize the prediction of the therapeutic activated partial thromboplastin time—age, ethnicity, gender, initial heparin dose, interval between initial heparin injection and first measurement of activated partial thromboplastin time, creatinine concentration, type of admission, and the aspartate aminotransferase to alanine aminotransferase ratio (AST/ALT ratio). These features contribute as a whole to patient outcomes, for example, creatinine in the blood is almost entirely filtered into the urine via glomerular filtration, and its concentration is stable under normal circumstances; therefore, creatinine concentration in the blood can be used as an indicator of renal function because it reflects the filtration function of glomeruli. Aspartate aminotransferase and alanine aminotransferase concentration levels in the blood are sensitive to hepatocellular damage, and their ratio is an important indicator of liver function. These features have been reported and discussed in another study [11], and many of the features exhibited statistically significant relationships with the first measurement of activated partial thromboplastin time after initial heparin dose.

Data Preprocessing

Patient data were preprocessed, and the features of interest were coded and normalized as variables. Missing values for some features were filled using the *k*–nearest neighbors algorithm which uses Euclidean distance to fill in missing values based on the values of its nearest neighbors in *k* dimensions.

Extreme values in data affect both the training and prediction processes. Normalization is needed when preprocessing continuous features; however, extreme values, though they may be few, negatively affect the output of normalization. Continuous features (age, heparin dose, creatinine value, and AST/ALT ratio) were manually verified to have *z* scores within the range of –3 to +3. According to the statistical definition of outliers [17], the normal range should be from $z=-3$ to $z=+3$; therefore, *z* scores outside of this range should be removed prior to normalization. Age data were found to be within the normal range; however, outliers were removed from initial heparin dose, creatinine concentration, and AST/ALT ratio data.

Model Training and Performance Tuning

The activated partial thromboplastin time value measured 4 to 6 hours after the initial heparin dose was classified using ternary classification into sub, normal, and supratherapeutic. The support vector machine, random forest, adaptive boosting, extreme gradient boosting, and shallow neural network algorithms were implemented and tested in this study.

A support vector machine is based on maximization of the margin (ie, the minimum distance from the separating hyperplane to the nearest data point) between 2 classes of data. A Gaussian kernel guarantees that classification is nonlinear. Adaptive boosting, extreme gradient boosting, and random forest methods are based upon the use of boosting as the method of learning. Boosting methods select features that are known to improve model predictive power, and thus simultaneously, to reduce dimensionality. Where typically sample features are the outputs of a weak classifier that has been applied to each sample, adaptive boosting trains different weak classifiers by changing the weight of the samples, and the weak class is combined into a weighted sum that represents the final output of the boosted classifier. Extreme gradient boosting is based on gradient boosting, a process in which the algorithm learns an ensemble of boosted trees and makes a careful tradeoff between the classification error and model complexity. Extreme gradient boosting has recently become dominant in the field of applied machine learning (for example, in Kaggle competitions for structured or tabular data) [18]. The random forest method grows multiple decision trees, each of which provides a classification. The forest chooses the final output by the classification that has the majority. Artificial neural networks are built of multiple layers of neurons; each neuron receives a number of input variables and passes on the results to neurons in the next layer. An artificial neural network can learn complex functions relating input to output variables and is able to deal with complex relationships between variables and functions. Our shallow neural network was built using TensorFlow (version 1.13.1).

Samples from subtherapeutic, normal therapeutic, and supratherapeutic data groups were included at a 1:1:1 ratio for training and validation of the ternary classification model. Each data set was divided into 80% training and cross-validation and 20% testing.

The best parameters for the support vector machine, random forest, adaptive boosting, and extreme gradient boosting algorithms were searched (GridSearch; scikitlearn package) and used to train the models. In the shallow neural network model, 2 hidden layers were used, and the number of neurons was set at 36/24 to reduce model complexity. To avoid overfitting, early stopping and regularization were needed. Dropout was also used since it is an effective method to avoiding overfitting and to improve robustness. The rectified linear unit activation function was chosen to increase nonlinearity [16,17,19]. The Adam optimizer was used in model training with an initial learning rate of 0.0015. We trained the model for 1500 epochs with the dropout rate set at 0.75. To validate the predictive performance of our models, 5-fold cross-validation was used on each.

Model Evaluation

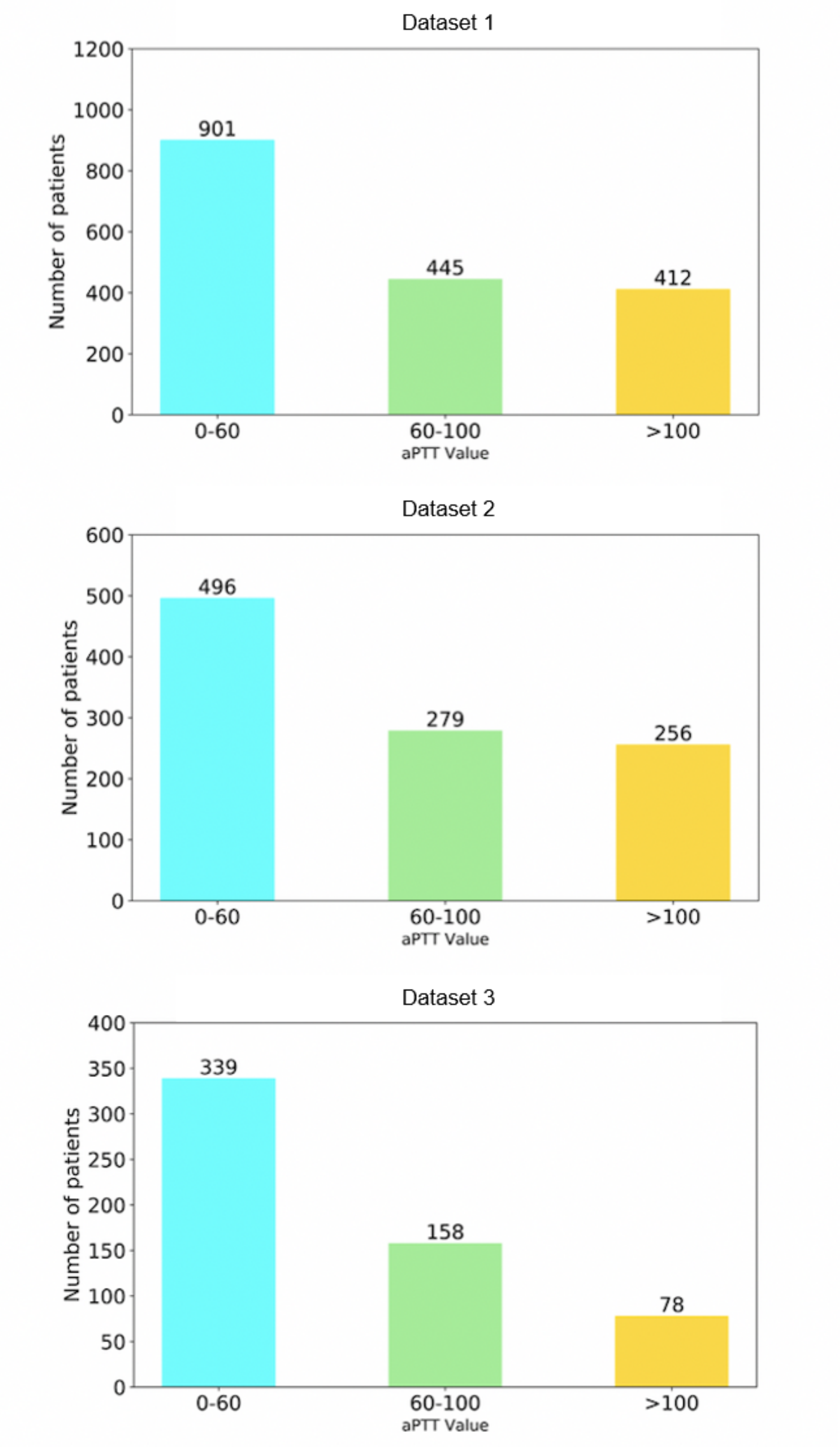
The following measures, *precision* = *true positive*/(*true positive* × *false positive*), *recall* = *true positive*/(*true positive* + *false negative*), *F1 score* = 2 × (*precision* × *recall*)/(*precision* + *recall*), and *accuracy* = (*true positive* + *true negative*)/(*true positive* + *true negative* + *false positive* + *false negative*), were used to evaluate the capability of our 3-class classification model [20]. For samples at a ratio of 1:1:1, the microaveraged precision, recall, and F1 score are all equal to the accuracy; therefore, we only compared the average accuracy and macroaveraged precision, recall, and F1 score to gauge the classification performances of these models.

Results

Activated Partial Thromboplastin Time Distribution in the Study Population

After removing outliers, we extracted data on intravenous push patients (data set 1, n=1758) and intravenous drip patients (data set 2, n=1031) who met our inclusion criteria from the MIMIC-III database and data on intravenous drip patients (data set 3, n=575) from the eICU database, respectively. In data set 1, 25.3% (445/1758) of patients had measured values of activated partial thromboplastin time within the normal therapeutic range, 51.3% (901/1758) had measured values of activated partial thromboplastin time within the subtherapeutic range, and 23.4% (412/1758) had measured values of activated partial thromboplastin time within the supratherapeutic range. In data set 2, 27.0% (279/1031), 48.1% (496/1031), and 24.9% (256/1031) of patients had measured values of activated partial thromboplastin time within the normal, subtherapeutic, and supratherapeutic ranges, respectively, as shown in Figure 1. In data set 3, 27.6% (158/575), 59.0% (339/575), and 13.6% (78/575) of patients had measured values of activated partial thromboplastin time within the normal, subtherapeutic, and supratherapeutic ranges, respectively.

Figure 1. Patient distribution of aPPT value after initial heparin dosing.



Summary Statistics of Selected Features

A descriptive summary of patient data in data set 1, 2, and 3 according to the therapeutic range of the first measurement of activated partial thromboplastin time after the initial heparin injection is shown in Table 1.

Table 1. Summary statistics of selected features.

Patient groups and features	Therapeutic range		
	Sub	Normal	Supra
Data set 1: MIMIC-III[a] intravenous push (N=1756), n	901	445	412
Age (years), mean (SD)	65.4 (14.6)	68.1 (15.1)	69.3 (14.2)
Initial heparin dose (units/hour), mean (SD)	907.0 (818.8)	1224.2 (1097.5)	1303.5 (908.4)
aPTT[b] (hours), mean (SD)	4.9 (0.6)	4.9 (0.6)	4.9 (0.6)
Ethnicity, n (%)			
White	639 (70.9)	311 (69.9)	291 (70.6)
Asian	11 (1.2)	5 (1.1)	7 (1.7)
Black	40 (4.4)	27 (6.1)	46 (11.2)
Hispanic/Latino	13 (1.4)	11 (2.5)	14 (3.4)
Others	198 (22.0)	91 (20.4)	54 (13.1)
Gender, n (%)			
Male	550 (61.0)	256 (57.5)	217 (52.7)
Female	351 (39.0)	189 (42.5)	195 (48.3)
Admission type, n (%)			
Elective	111 (12.3)	26 (5.8)	15 (3.6)
Emergency	768 (79.8)	398(89.4)	388 (94.2)
Urgent	32 (3.6)	21 (4.7)	9 (2.2)
Data set 2: MIMIC-III intravenous drip (N=1031), n	496	279	256
Age (years), mean (SD)	64.9 (15.4)	68.6 (15.2)	70.1 (14.8)
Initial heparin dose (units/hour), mean (SD)	969.4 (398.3)	1148.7 (395.8)	1229.8 (495.3)
aPTT (hours), mean (SD)	5.0 (0.6)	4.9 (0.5)	5.0 (0.6)
Ethnicity, n (%)			
White	353 (71.2)	208 (74.6)	179 (70.0)
Asian	9 (1.8)	9 (3.2)	10 (3.9)
Black	46 (9.3)	29 (10.4)	42 (16.4)
Hispanic/Latino	12 (2.4)	7 (2.5)	9 (3.5)
Others	76 (15.3)	26 (9.3)	15 (6.2)
Gender, n (%)			
Male	312 (62.9)	163 (58.4)	132 (51.6)
Female	184 (37.1)	116 (41.6)	124 (48.4)
Admission type, n (%)			
Elective	59 (11.9)	25 (9.0)	8 (3.1)
Emergency	436 (87.9)	250 (89.6)	245 (95.7)
Urgent	1 (0.2)	4 (1.4)	3 (1.2)
Data set 3: eICU[d] intravenous drip (N=575), n	339	158	78
Age (years), mean (SD)	64.8 (13.9)	69.0 (14.4)	73.1 (12.3)
Initial heparin dose (units/hour), mean (SD)	1005.7 (892.6)	973.5 (519.3)	950.4 (539.4)
aPTT (hours), mean (SD)	5.2 (0.6)	5.2 (0.6)	5.2 (0.6)
Ethnicity, n (%)			
White	244 (72.0)	106 (67.1)	46 (59.0)

Patient groups and features	Therapeutic range		
	Sub	Normal	Supra
Asian	4 (1.2)	2 (1.3)	2 (2.6)
Black	30 (8.8)	20 (12.7)	9 (11.5)
Hispanic/Latino	22 (6.5)	19 (12.0)	12 (15.4)
Others	39 (11.5)	11 (7.0)	9 (11.5)
Gender, n (%)			
Male	217 (64.3)	99 (62.7)	37 (47.4)
Female	122 (35.7)	59 (37.3)	41 (52.6)
Creatinine (mg/dL), mean (SD)	1.7 (1.7)	2.0 (2.1)	2.0 (1.5)
AST/ALT[c], mean (SD)	1.5 (1.2)	1.7 (1.3)	1.5(1.1)

[a] Multiparameter Intelligent Monitoring In Intensive Care III database.

[b] First measurement of activated partial thromboplastin time.

[c] AST/ALT: aspartate aminotransferase ratio/alanine aminotransferase.

[d] eICU: e–Intensive Care Unit database.

Data Preprocessing Results

Outliers were removed for 3 features: heparin dose, creatinine value, and AST/ALT ratio. The statistical outliers are shown in Multimedia Appendix 1. Not all patients had a complete set of clinical data, for example, 154 patients were missing AST/ALT ratios, accounting for 8.76% of intravenous push patients (Multimedia Appendix 2). An algorithm (*k* nearest neighbors) was used to fill in the missing values. Since filled values accounting for up to 40% have been reported to be appropriate [21], we considered the effect of filled features on the activated partial thromboplastin time as reasonable.

Model Performance Results

To eliminate category imbalances, we randomly selected 400 samples for each therapeutic state in data set 1, 250 samples for each therapeutic state in data set 2, and 120 samples for each therapeutic state in data set 3. For subtherapeutic and normal therapeutic classes, general downsampling was used to reduce the number of samples, while for the supratherapeutic class we used upsampling to increase the number of samples to 120; therefore, experiments used 1200 samples from data set 1, 750 samples from data set 2, and 360 samples from data set 3. Model performance results are shown in Table 2.

The F1 score provides a comprehensive evaluation of the model. As listed in Multimedia Appendix 3, extreme gradient boosting achieved the second best F1 scores (77.58%, 73.94%, and 78.85% for data set 1, 2, and 3, respectively), second only to those of the shallow neural network (87.26%, 85.98% and 87.55% for data set 1, 2, and 3, respectively). The adaptive boosting model also performed very well in all 3 data sets (72.80%, 81.67%, and 77.65% for data set 1, 2, and 3, respectively), with scores close to those of extreme gradient boosting (77.58%, 73.94%, and 78.85% for data set 1, 2, and 3, respectively). The random forest performed slightly worse (68.20%, 73.15%, and 65.59% for data set 1, 2, and 3, respectively) than the other 4 models. The confusion matrices of all 5 models are shown in Multimedia Appendix 4. In further experiments, the random forest still performed better than other models that were not discussed herein, such as the Naïve Bayes, logistic regression, *k* nearest neighbors, and decision tree, as shown in Multimedia Appendix 3.

Table 2. Macroaveraged scores for the machine learning algorithms.

Models	Precision, %	Recall, %	F1 score, %	Accuracy, %
Data set 1: MIMIC-III[a] (intravenous push patients)				
Random forest	68.96	68.75	68.70	68.75
Adaptive boosting	74.37	72.92	72.80	72.92
Support vector machine	85.19	73.33	73.79	73.33
Extreme gradient boosting	79.27	76.25	77.58	76.25
Shallow neural network	88.05	86.67	87.26	86.67
Data set 2: MIMIC-III (intravenous drip patients)				
Random forest	66.71	65.33	65.06	65.33
Adaptive boosting	77.29	77.33	77.30	77.33
Support vector machine	84.59	71.33	71.71	71.33
Extreme gradient boosting	77.45	77.33	77.38	77.33
Shallow neural network	85.99	86.00	85.98	86.00
Data set 3: eICU[b] (intravenous drip patients)				
Random forest	66.77	66.56	65.59	68.06
Adaptive boosting	78.03	77.78	77.65	77.78
Support vector machine	84.74	76.39	76.19	76.39
Extreme gradient boosting	79.16	79.17	78.85	79.17
Shallow neural network	87.80	87.50	87.55	87.50

[a] Multiparameter Intelligent Monitoring In Intensive Care III database.

[b] eICU: e–Intensive Care Unit database.

In the subtherapeutic class, adaptive boosting achieved the highest precision in data set 1 (84.48%) while the neural network model achieved highest in the other data sets (data set 2: 83.67%; data set 3: 83.33%). The support vector machine achieved the highest recall in all 3 data sets (data set 1: 100%; data set 2: 100%; data set 3: 95.83%). In the normal therapeutic class, the support vector machine with the Gaussian kernel achieved 100% precision in all 3 data sets. The shallow neural network achieved the highest recall (data set 1: 85.00%; data set 2: 86.00%; data set 3: 91.67%). In the supratherapeutic class, the support vector machine achieved the highest precision (data set 1: 100%; data set 2: 100%; data set 3: 95.24%); however, recall of the support vector machine was not very high (data set 1: 57.50%; data set 2: 58.00%; data set 3: 83.33%). The shallow neural network achieved the best recall in all 3 data sets (data set 1: 100%; data set 2: 100%; data set 3: 95.83%). Considering the comprehensive performance which is best evaluated by F1 score, the shallow neural network achieved the best F1 score in all 3 patient groups: subtherapeutic (data set 1: 79.35%; data set 2: 83.67%; data set 3: 83.33%), normal therapeutic (data set 1: 93.15%; data set 2: 87.76%; data set 3: 84.62%), and supratherapeutic (data set 1: 88.00%; data set 2: 86.54%; data set 3: 95.45%) therapeutic ranges. Additional results are listed in Table 3, Table 4, and Table 5.

Table 3. Model performance for subtherapeutic.

Models	Precision, %	Recall, %	F1 score, %
Data set 1: MIMIC-III[a] (intravenous push patients)			
Random forest	67.61	60.00	63.58
Adaptive boosting	84.48	61.25	71.01
Support vector machine	55.56	100	71.43
Extreme gradient boosting	74.32	68.75	71.43
Shallow neural network	79.35	91.25	84.89
Data set 2: MIMIC-III (intravenous drip patients)			
Random forest	58.82	80.00	67.80
Adaptive boosting	78.43	80.00	79.21
Support vector machine	53.76	100	69.93
Extreme gradient boosting	74.32	68.75	71.43
Shallow neural network	83.67	82.00	82.83
Data set 3: eICU[b] (intravenous drip patients)			
Random forest	66.67	58.33	62.22
Adaptive boosting	77.27	70.83	73.91
Support vector machine	58.97	95.83	73.02
Extreme gradient boosting	76.00	79.17	77.55
Shallow neural network	83.33	83.33	83.33

[a] Multiparameter Intelligent Monitoring In Intensive Care III database.

[b] eICU: e–Intensive Care Unit database.

Table 4. Model performance for normal therapeutic.

Models	Precision, %	Recall, %	F1 score, %
Data set 1: MIMIC-III[a] (intravenous push patients)			
Random forest	63.64	84.00	72.41
Adaptive boosting	71.26	77.50	74.25
Support vector machine	100	62.50	76.92
Extreme gradient boosting	78.72	74.00	76.29
Shallow neural network	93.15	85.00	88.89
Data set 2: MIMIC-III (intravenous drip patients)			
Random forest	73.81	62.00	67.39
Adaptive boosting	78.43	80.00	79.21
Support vector machine	100	56.00	71.79
Extreme gradient boosting	72.55	74.00	73.27
Shallow neural network	87.76	86.00	86.87
Data set 3: eICU[b] (intravenous drip patients)			
Random forest	70.00	65.00	61.90
Adaptive boosting	81.82	53.85	78.26
Support vector machine	100	50.00	66.67
Extreme gradient boosting	80.00	66.67	72.73
Shallow neural network	84.62	91.67	87.50

[a] Multiparameter Intelligent Monitoring In Intensive Care III database.

[b] eICU: e–Intensive Care Unit database.

Table 5. Model performance for supratherapeutic.

Models	Precision, %	Recall, %	F1 score, %
Data set 1: MIMIC-III[a] (intravenous push patients)			
Random forest	75.95	75.00	75.47
Adaptive boosting	67.37	80.00	73.14
Support vector machine	100	57.50	73.02
Extreme gradient boosting	80.77	78.75	79.75
Shallow neural network	88.00	82.50	85.16
Data set 2: MIMIC-III (intravenous drip patients)			
Random forest	67.50	54.00	60.00
Adaptive boosting	75.00	72.00	73.47
Support vector machine	100	58.00	73.42
Extreme gradient boosting	76.47	78.00	77.23
Shallow neural network	86.54	90.00	88.24
Data set 3: eICU[b] (intravenous drip patients)			
Random forest	63.64	87.50	73.68
Adaptive boosting	75.00	87.50	80.77
Support vector machine	95.24	83.33	88.89
Extreme gradient boosting	81.48	91.67	86.27
Shallow neural network	95.45	87.50	91.30

[a] Multiparameter Intelligent Monitoring In Intensive Care III database.

[b] eICU: e–Intensive Care Unit database.

Discussion

Principal Results

In our experiments, the neural network achieved the highest scores for all evaluation metrics. The neural network model uses multiple layers to progressively extract higher level features from the raw data which might be the reason that the neural network is able to learn some unknown features that help to provide a better classification of normal therapeutic activated partial thromboplastin time. Since different features may be correlated (such as the creatinine value and aspartate aminotransferase), linear classification models are not appropriate. Random forest, adaptive boosting, and extreme gradient boosting are ensemble learning methods. By integrating weak classifiers, classification performance was greatly improved. The support vector machine with Gaussian kernel is a widely used and powerful classifier. Gaussian kernels ensure that the classifier is nonlinear, which suited the characteristics of our data, and the method was able to demonstrate high performance; however, the neural network model was able to take into account complex relationships between the variables with complex functions. Among the methods tested, the shallow neural network performed the best. The shallow neural network achieved performance approximately 10% higher than that of the other algorithms for each metric (precision, recall, F1 score, and accuracy) in intravenous push cases (data set 1) and achieved performance approximately 9% higher than that of the other algorithm metrics in intravenous drip cases (data set 2 and data set 3). Extreme gradient boosting, adaptive boosting, and the support vector machine were the models that subperformed to the shallow neural network although their scores were, nevertheless, all above 70%. The random forest model demonstrated the worst performance.

As a result of its relative high accuracy, this shallow neural network model should be able to recommend doses better than the heparin dosage guidelines which only take patient weight into account.

In clinical practice, intravenous push and intravenous drip are both common delivery routes for heparin. Intravenous push heparin is always used to rescue critical patients who require timely intervention to decrease coagulation, while intravenous drip heparin is used a long-term medication to prevent thrombosis or embolic disease. These 2 administration routes have different clinical significance; therefore, we separated the patient groups from the 2 databases into 3 data sets to verify whether they would have different model predictions. The results suggested that model prediction performance was comparable among the 3 data sets, which gave us insight into the stability and suggests the model is stable regardless of administration routes or data source.

Strengths

Since the range of normal therapeutic activated partial thromboplastin time varies in different institutions, our shallow neural network model can be adapted to different heparin administration guidelines by adjusting the parameters.

Furthermore, the model can also be applied to other drug dosage optimization problems after retraining. When treating a patient, a dose of heparin can be recommended that maximizes the normal therapeutic probability. The future application of the model prediction has the potential to enhance patient safety, minimize the risk of bleeding or a thromboembolic event, reduce medical costs, and improve the efficiency of clinicians.

Limitations

One challenge of our study was to identify the features that affect heparin doses. First, balancing both discrete features and continuous features and their relative importance would have enhanced model training performance and feature utilization but was not performed in this study. Second, different features may have been correlated, since they all contribute to the comprehensive conditions of patients; therefore, determining the intrinsic relationships would have further improved model performance. Model optimization and verification using different intensive care unit databases will be performed in future research. Drug interactions with heparin and the accumulated effects are usually not taken into account since the half-time of heparin is too short to affect the 4 to 6–hour interval that was monitored. A more precise neural network structure was not used; the next step would be to explore the intrinsic relationships between features and further validate the model results using additional clinical data sets. Since this study was conducted in a nonclinical setting, it will be further refined as it is used in practice.

Comparison With Prior Work

It is difficult to obtain personalized rather than broad normative data to determine drug dosage in intensive care units. Heparin dose is commonly determined based solely upon body weight, which is measured or estimated when patients arrive at the intensive care unit. Here, we distinguished 2 drug delivery routes to provide more detailed advice and choices for clinicians. The overall prediction accuracies for the 3 data sets were 88.00%, 86.00%, and 87.50%. Both delivery routes in the MIMIC-III retrospective data showed proportions of patients with activated partial thromboplastin times that were 3-fold higher than those with normal therapeutic activated partial thromboplastin times (25.3% for intravenous push patients and 27.0% for intravenous drip patients), and higher than those reported in previous studies [11,12] for the multivariate logistic regression (volume under the surface=0.48) and multinomial logistic regression (accuracy=60%). Statistical results were consistent with those from previous reports. Advanced age and gender (female) were reported to be associated with supratherapeutic activated partial thromboplastin time [9,10], as well as a high initial heparin dose, a high AST/ALT ratio, and emergency admission-type.

Conclusions

The study aimed to provide support to predict heparin treatment outcomes and recommend optimal heparin dosing to clinicians. Data-driven machine learning methods were used to predict the probabilities of subtherapeutic, normal therapeutic, and supratherapeutic activated partial thromboplastin time. After comparing different models, we recommend the adoption of a support system comprising a shallow neural network with parameter adjustability. The results of this study provide new insights into personalized medication optimization and demonstrate the feasibility of applying the model in different medical institutions.

Acknowledgments

This study was supported by the National Key Research & Development Program of China (project 2018YFC0116905).

Authors' Contributions

YL, WZ, and NH are corresponding authors and take responsibility for the integrity of the work as a whole. LS, CL, DL, and JH contributed equally as co–first authors and were responsible for study conception and design. JH, FZ, and HJ were responsible for data cleaning and algorithm implementation. DL, MG, and HW were responsible for data analysis and explanation of results. LS, CL, NH, and JH drafted the manuscript. All authors revised the manuscript for important intellectual content.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Outliers preprocessing.
[DOCX File , 41 KB - medinform_v8i6e17648_app1.docx]

Multimedia Appendix 2
Missing data imputation.
[DOCX File , 13 KB - medinform_v8i6e17648_app2.docx]

Multimedia Appendix 3
Macroaveraged scores of different algorithms on 3 different datasets.
[DOCX File , 18 KB - medinform_v8i6e17648_app3.docx]

Multimedia Appendix 4
Confusion matrix.
[DOCX File , 17 KB - medinform_v8i6e17648_app4.docx]

References

1. Williams CN, Bratton SL, Hirshberg EL. Computerized decision support in adult and pediatric critical care. World J Crit Care Med 2013 Dec 04;2(4):21-28 [FREE Full text] [doi: 10.5492/wjccm.v2.i4.21] [Medline: 24701413]
2. Pirracchio R, Cohen MJ, Malenica I, Cohen J, Chambaz A, Cannesson M, ACTERREA Research Group. Big data and targeted machine learning in action to assist medical decision in the ICU. Anaesth Crit Care Pain Med 2019 Aug;38(4):377-384 [FREE Full text] [doi: 10.1016/j.accpm.2018.09.008] [Medline: 30339893]
3. Raschke RA, Reilly BM, Guidry JR, Fontana JR, Srinivas S. The weight-based heparin dosing nomogram compared with a "standard care" nomogram. A randomized controlled trial. Ann Intern Med 1993 Dec 01;119(9):874-881. [doi: 10.7326/0003-4819-119-9-199311010-00002] [Medline: 8214998]
4. Raschke R, Gollihare B, Peirce JC. The effectiveness of implementing the weight-based heparin nomogram as a practice guideline. Arch Intern Med 1996;156(15):1645-1649. [Medline: 8694662]
5. Spruill W, Wade W, Huckaby W, Leslie RB. Achievement of anticoagulation by using a weight-based heparin dosing protocol for obese and nonobese patients. Am J Health Syst Pharm 2001 Dec 15;58(22):2143-2146. [doi: 10.1093/ajhp/58.22.2143] [Medline: 11760916]
6. Gerlach A, Folino J, Morris BN, Murphy CV, Stawicki SP, Cook CH. Comparison of heparin dosing based on actual body weight in non-obese, obese and morbidly obese critically ill patients. Int J Crit Illn Inj Sci 2013 Jul;3(3):195-199 [FREE Full text] [doi: 10.4103/2229-5151.119200] [Medline: 24404457]
7. Hohner E, Kruer R, Gilmore V, Streiff M, Gibbs H. Unfractionated heparin dosing for therapeutic anticoagulation in critically ill obese adults. J Crit Care 2015 May;30(2):395-399. [doi: 10.1016/j.jcrc.2014.11.020] [Medline: 25534987]
8. Fan J, John B, Tesdal E. Evaluation of heparin dosing based on adjusted body weight in obese patients. Am J Health Syst Pharm 2016 Oct 01;73(19):1512-1522. [doi: 10.2146/ajhp150388] [Medline: 27646813]
9. Melloni C, Alexander KP, Chen AY, Newby LK, Roe MT, Allen LaPointe NM, CRUSADE Investigators. Unfractionated heparin dosing and risk of major bleeding in non-ST-segment elevation acute coronary syndromes. Am Heart J 2008 Aug;156(2):209-215 [FREE Full text] [doi: 10.1016/j.ahj.2008.03.023] [Medline: 18657648]
10. Lee M, Wali A, Menon V, Berkowitz S, Thompson T, Califf R, et al. The determinants of activated partial thromboplastin time, relation of activated partial thromboplastin time to clinical outcomes, and optimal dosing regimens for heparin treated patients with acute coronary syndromes: a review of GUSTO-IIb. J Thromb Thrombolysis 2002 Oct;14(2):91-101. [doi: 10.1023/a:1023235926825] [Medline: 12714828]
11. Ghassemi MM, Richter SE, Eche IM, Chen TW, Danziger J, Celi LA. A data-driven approach to optimized medication dosing: a focus on heparin. Intensive Care Med 2014 Oct 5;40(9):1332-1339 [FREE Full text] [doi: 10.1007/s00134-014-3406-5] [Medline: 25091788]
12. Ghassemi M, Alhanai T, Westover M, Mark R, Nemati S. Personalized Medication Dosing Using Volatile Data Streams. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018 Jun 20 Presented at: Personalized medication dosing using volatile data streams. Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence; 2018; the Hilton New Orleans Riverside, New Orleans, Louisiana, USA URL: https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/17234
13. Grzymala-Lubanski B, Själander S, Renlund H, Svensson PJ, Själander A. Computer aided warfarin dosing in the Swedish national quality registry AuriculA - Algorithmic suggestions are performing better than manually changed doses. Thromb Res 2013 Mar;131(2):130-134. [doi: 10.1016/j.thromres.2012.11.016] [Medline: 23232091]
14. Jacobs M. Personalized anticoagulant management using reinforcement learning. ELECTRONIC THESES AND DISSERTATIONS 2014 May:670 [FREE Full text] [doi: 10.18297/etd/670]
15. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3(1):160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
16. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5(1):180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
17. Cousineau D, Chartier S. Outliers detection and treatment: a review. Int. j. psychol. res 2010 Jun 30;3(1):58-67. [doi: 10.21500/20112084.844]
18. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international Conference on Knowledge Discovery and Data Mining. 2016 Aug Presented at: Knowledge Discovery and Data Mining; August 2016; San Francisco p. 785-794. [doi: 10.1145/2939672.2939785]
19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems 25. 2012 Presented at: International Conference on Neural Information Processing Systems; 2012; Lake Tahoe, Nevada, United States.

20. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management 2009 Jul;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
21. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition 2008 Dec;41(12):3692-3705 [FREE Full text] [doi: 10.1016/j.patcog.2008.05.019]

Abbreviations

ALT: alanine aminotransferase
AST: aspartate aminotransferase
eICU: e–Intensive Care Unit
MIMIC-III: Multiparameter Intelligent Monitoring in Intensive Care III (or Medical Information Mart for Intensive Care III)

Edited by T Hao, B Tang; submitted 31.12.19; peer-reviewed by Z Zhang, S Ayabakan; comments to author 23.02.20; revised version received 18.04.20; accepted 19.04.20; published 22.06.20.

Please cite as:
Su L, Liu C, Li D, He J, Zheng F, Jiang H, Wang H, Gong M, Hong N, Zhu W, Long Y
Toward Optimal Heparin Dosing by Comparing Multiple Machine Learning Methods: Retrospective Study
JMIR Med Inform 2020;8(6):e17648
URL: http://medinform.jmir.org/2020/6/e17648/
doi:10.2196/17648
PMID:32568089

Original Paper

Automatic Construction of a Depression-Domain Lexicon Based on Microblogs: Text Mining Study

Genghao Li[1], MSF; Bing Li[1], PhD; Langlin Huang[1], BM; Sibing Hou[2], MAFN

[1]School of Information Technology & Management, University of International Business and Economics, Beijing, China

[2]Graduate School of Art and Science, Columbia University, New York, NY, United States

Corresponding Author:
Bing Li, PhD
School of Information Technology & Management
University of International Business and Economics
Chaoyang District, Huixin East Street
Beijing, 100029
China
Phone: 86 1 343 978 8086
Email: 01630@uibe.edu.cn

Abstract

Background: According to a World Health Organization report in 2017, there was almost one patient with depression among every 20 people in China. However, the diagnosis of depression is usually difficult in terms of clinical detection owing to slow observation, high cost, and patient resistance. Meanwhile, with the rapid emergence of social networking sites, people tend to share their daily life and disclose inner feelings online frequently, making it possible to effectively identify mental conditions using the rich text information. There are many achievements regarding an English web-based corpus, but for research in China so far, the extraction of language features from web-related depression signals is still in a relatively primary stage.

Objective: The purpose of this study was to propose an effective approach for constructing a depression-domain lexicon. This lexicon will contain language features that could help identify social media users who potentially have depression. Our study also compared the performance of detection with and without our lexicon.

Methods: We autoconstructed a depression-domain lexicon using Word2Vec, a semantic relationship graph, and the label propagation algorithm. These two methods combined performed well in a specific corpus during construction. The lexicon was obtained based on 111,052 Weibo microblogs from 1868 users who were depressed or nondepressed. During depression detection, we considered six features, and we used five classification methods to test the detection performance.

Results: The experiment results showed that in terms of the F1 value, our autoconstruction method performed 1% to 6% better than baseline approaches and was more effective and steadier. When applied to detection models like logistic regression and support vector machine, our lexicon helped the models outperform by 2% to 9% and was able to improve the final accuracy of potential depression detection.

Conclusions: Our depression-domain lexicon was proven to be a meaningful input for classification algorithms, providing linguistic insights on the depressive status of test subjects. We believe that this lexicon will enhance early depression detection in people on social media. Future work will need to be carried out on a larger corpus and with more complex methods.

(JMIR Med Inform 2020;8(6):e17650) doi:10.2196/17650

KEYWORDS

depression detection; depression diagnosis; social media; automatic construction; domain-specific lexicon; depression lexicon; label propagation

Introduction

Background

Depression, one of the major reasons for suicide in recent years, is a severe mental disorder characterized by persisting low mood states in the affected person. It is expected to be the largest contributor to disease burden worldwide by 2030, especially in China with a high-pressure lifestyle. According to a World Health Organization (WHO) report in 2017 [1], China had more than 54 million people with depression, which means that there was almost one patient with depression among every 20 people. In addition, a national estimation based on China’s 2012 census data shows that with an adult population size of 1.04 billion, an estimated 258.41 million adults (24.79%) are at increased risk of depressive symptoms [2]. It has been reported that the suicide rate among patients with depression is more than 20 times that of the general population, and patients with depression account for more than half of those who have committed suicide [3].

Diagnosis of potential depression in an early stage can provide more opportunities for those affected to receive appropriate treatment and overcome the disease. However, owing to the lack of mental health knowledge, the lack of regular counseling, and the fact that mental health diseases are greatly different from physical diseases as there is no pain, many patients with depression do not recognize it. Although some know a little about depression, they are often reluctant to seek professional help because of a sense of shame [4].

The traditional clinical diagnosis of depression mainly relies on standardized assessments, which are highly accurate but have limitations in detection efficiency [5]. The medical diagnosis requires not only filling in a depression assessment scale, such as the Self-rating Depression Scale, but also a one-to-one interview and long-term observation [6], which involve high costs. Patients tend to remain undetected until the disease presents obvious symptoms, which also means that the optimal treatment period has passed [7]. The whole diagnosis process is highly passive, as doctors have to wait for patients to knock on their door.

Things are changing with the development of social media. Nowadays, many methods combining machine learning algorithms and text mining techniques have been developed to diagnose potential depression in an early stage [8-13]. Compared with traditional approaches, these methods have been proven to be effective and inexpensive, and have been shown to reduce limitations and assist in clinical diagnosis in a more flexible way. At the same time, people are used to disclosing their inner feelings on social media. The huge corpus provides abundant text describing things like sadness, exhaustion, and breakdown, which have the potential to reflect depression. Hundreds of millions of people in third or fourth tier cities and poor mountainous areas in China have little chance to disclose their mental conditions directly to experts, but they can provide their accounts and apply for social media methods. Experts can then intervene and conduct more targeted treatments for users who are potentially depressed. Another scenario involves teenagers on campus, and teachers can pay more attention to the actual mental status of students who are potentially depressed with the help of forums and other web-based text. It is thus feasible to detect users’ depressive mental states on a large scale on social media, and this provides convenience for expert assessment.

Actually, when coping with textual depression data, word-based features like frequency and embedding are commonly used and a domain lexicon might be valuable to understand the author of the text [14]. Many research studies have achieved a lot in terms of a depression lexicon, which is mainly in English [9,12,13]. In China, research about web-related depression detection is just getting started, and we did not find any domain lexicon research about depression in a public study. It would not be proper to translate an English lexicon directly owing to cultural differences. Thus, a depression-domain lexicon in Chinese is needed.

In this paper, based on a well-labeled depression data set on Weibo, which is one of the largest Chinese user-generated content platforms, we constructed a depression-domain lexicon containing more than 2000 words. This lexicon can be used to assist in the early diagnosis of depression. We crawled more than 144,000 microblog tweets of nearly 2000 users within a time span of 16 months to obtain depressed and nondepressed data sets. Some manual screening was implemented to remove “fake” depression microblogs from the data sets, which is clarified in the “Data Preprocessing” subsection. We extracted 80 words as seeds and then built a semantic association graph with the similarities between the seeds and candidate words and utilized the label propagation algorithm (LPA) to automatically mark new words in the graph. The LPA is a good method in such a construction, which has been further explained in the “Related Work” subsection. We then tested the effectiveness of this method and compared it with some baseline approaches. We found that this autoconstruction of a depression-domain lexicon performed the best and had the most stable performance when parameters changed. For further research, this lexicon was used as an input for machine learning algorithms, providing insights into the depressive status of test subjects, so as to improve detection accuracy. According to our research, the detection models with lexicon features outperformed the models without lexicon features by 2% to 9% in terms of evaluation scores.

The main contributions are as follows: (1) We extracted a set of depressive words and constructed our domain lexicon in Chinese, which is a good contribution to web-related depression signal detection, to assist in identifying users who have the potential to experience depression in an early period. We applied an efficient semisupervised automatic construction method in the depression domain. The lexicon was proven to be meaningful in several detecting classification models in our study; (2) We constructed a benchmark depression data set (some of the data were used to construct the lexicon [our main research objective] and the other data were used in the detection test) based on microblogs, which could assist in further depression detection, diagnosis, and analysis. Meanwhile, we released the data set and lexicon together [15] to facilitate future web-related depression diagnosis.

Related Work About the Traditional Approach for Depression Detection

For decades, there have been many ways to detect depression. Beck [16] created the original Beck Depression Inventory for a quick self-testing measure that can briefly assess recent depression symptoms. Thereafter, Beck et al [17] updated the approach to Beck Depression Inventory II that can assess the severity of self-reported depression symptoms by paper or electronic format. Radloff [18] developed the Center for Epidemiologic Studies Depression Scale, which focuses more on the individual’s emotional experience and less on the physical condition. Some other popular scales are the Zung Self-rating Depression Scale [19] and Hamilton Depression Rating Scale [20].

Since the 21st century, new scales are continuously being improved. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) is a standard classification of mental disorders used by mental health professionals, which was improved by Hu [21]. In China, the Chinese Classification and Diagnosis of Mental Diseases 3rd edition (CCMD-3) is a standard for diagnosis.

Overall, traditional ways of depression detection have been highly validated and accepted in the real world for decades. However, they mainly rely on the scores of scales or questionnaires, face-to-face interviews, and self-reports, and often require a lot of labor and time [6]. The new trend might be related to big data that are timely, rich, and easily accessible on social networking sites like Facebook, Twitter, and Weibo. These web-based methods can assist in large-scale early detection, and experts can further conduct more precise diagnosis and treatment.

Related Work About Depression Detection on Social Media

In recent years, with abundant data on social media, some researchers are attempting to detect depression by leveraging web-based data. Park et al [8] explored the use of depressive language from Twitter users and concluded that social networks can provide meaningful data for capturing depressive moods. Choudhury et al [9] were the first to diagnose and predict depression via social media by extracting several features. Hasan et al [10] proposed a new method with the circumplex model to classify Twitter messages as depressed, happy, or other emotions. Resnik et al [11] explored the use of supervised topic models in an analysis of linguistic signals for detecting depression. During such research, word-based features are of great importance on social media [14].

Word-based features could be shown in a lexicon. Tsugawa et al [12] utilized positive and negative sentiment words in a lexicon for recognizing depression. Choudhury et al [9] used semantic orientation pointwise mutual information (SO-PMI) and term frequency-inverse documentation frequency (TF-IDF) to extract a depression lexicon from “mental health” in Yahoo! Answers and set the Wikipedia page on “list of antidepressants” as antidepressant words. Most recently, Guangyao et al [13] employed Word2Vec (W2V) to extract words of antidepressants and depression symptoms from Twitter as a domain-specific lexicon.

Many previous studies achieved a lot with regard to a depression lexicon, which can greatly help in diagnosis; however, most of the words are in English. It is not proper to use the translated version of an English lexicon to detect depression in a Chinese corpus because of cultural differences. In addition, only PMI (mainly about co-occurrence frequency) and W2V (word embeddings) techniques cannot keep up with today’s semantic developments. We can see the feasibility of detecting depression on social media with a lexicon, and more efforts are needed to construct a better Chinese depression-domain lexicon.

Related Work About Research on Construction of a Domain Lexicon

Many methods have been used to efficiently construct a domain lexicon. Das et al [22] and Krestel and Siersdorfer [23] used SO-PMI as a useful tool for emotion lexicon construction. Yu and Dredze [24], Tixier et al [25], and Zhengyu [26] leveraged and improved the W2V method to construct a domain lexicon. Chao et al [27] proposed a semisupervised sentiment orientation classification algorithm based on W2V (SO-W2V) and obtained a lexicon in different areas efficiently. The PMI method focuses on the co-occurrence frequency between words but ignores the context. However, W2V considers context with word embeddings but in a relatively simple way compared with the LPA shown below.

The LPA, which was first proposed by Zhu and Ghahramaniy [28], plays an important role in lexicon autoconstruction with semisupervised methods. Researchers [29,30] used the LPA starting with several labeled seed words to expand a lexicon for polarity classification. Tai and Kao [31] built a framework to automatically generate a lexicon by combining PMI and the LPA. Hamilton et al [32] applied a label propagation framework with domain-specific word embeddings to construct accurate domain-specific lexicons. A new method combining W2V and LPA was adopted by Giulianelli [33] and Pu et al [34], and it performed much better than previous methods. In this way, the relationships between words and specific domain contexts are considered.

Data Collection

In order to build a depression-domain lexicon for further detection via social media, we constructed two data sets of users with depression and without depression based on data from Weibo microblogs, which is very popular in China. Weibo has 462 million monthly active users according to a report in 2018 [35], and it is the most popular social media website in China. Equivalent with Twitter, people are getting used to sharing their ideas and moods on Weibo.

In light of the fact that depression is a long-standing illness, the text of users should not be collected from only one microblog. Thus, our data sets contained all Weibo microblogs within a year published by the same users. In addition, personal profile information like comments, number of follows, and number of followers was also included.

Depressed Data Set D1

Based on Weibo microblogs from January 2017 to April 2018, we used the keywords “I’m diagnosed with depression” [13,36,37] to construct a depressed data set $D1$. In this way, we finally identified 965 users with depression and 58,265 microblogs (Table 1).

Table 1. Details of the collected data sets from Weibo microblogs.

Data set	Users	Total posts	Mean	Standard deviation	Skewness	Kurtosis	Time span
Depressed data set *D*1	965	58,265	60.374	31.327	−0.451	1.788	January 2017-April 2018 (16 months)
Nondepressed data set *D*2	903	52,787	63.697	30.086	−0.615	2.066	January 2017-April 2018 (16 months)

Nondepressed Data Set D2

If a user never posted any text with a depression-related word like "depress," the user was labeled as nondepressed. In this way, we constructed a nondepressed data set *D*2. To match *D*1, we selected a similar number of microblogs (one user without depression can have up to 100 posts) under the same time span. In this way, we identified 903 users without depression and 52,787 microblogs (Table 1).

Data Preprocessing

Before the experiment, we found that there were some unrelated microblogs, irregular words, and emoji in our data sets. These noisy texts can affect the accuracy of our model, so we adopted the following preprocessing procedures: (1) *Emoji processing*. Emoji and emoticons are common in social media. However, they can cause some unexpected troubles like encoding problems during algorithm running and text analysis, so we removed emoji. We will take them into account separately in further research; (2) *Unrelatedmicroblog processing*. In addition to depression-domain microblogs that we focused on mostly, many users posted plenty of daily microblogs, including red packets snatching, game sharing, advertisements, etc. In addition, some "fake" depression microblogs like depression scientific articles and content talking about friends with depression, instead of users, are also useless and can be misleading. By manual screening, we obtained a list of unrelated keywords in daily microblogs and "fake" microblogs, and then, we removed them all with regular expression; (3) *Irregular words preprocessing*. New words keep appearing, and language habits are quite different on the internet. These cause trouble during text analysis. Therefore, we added a general dictionary of internet words.

Methods

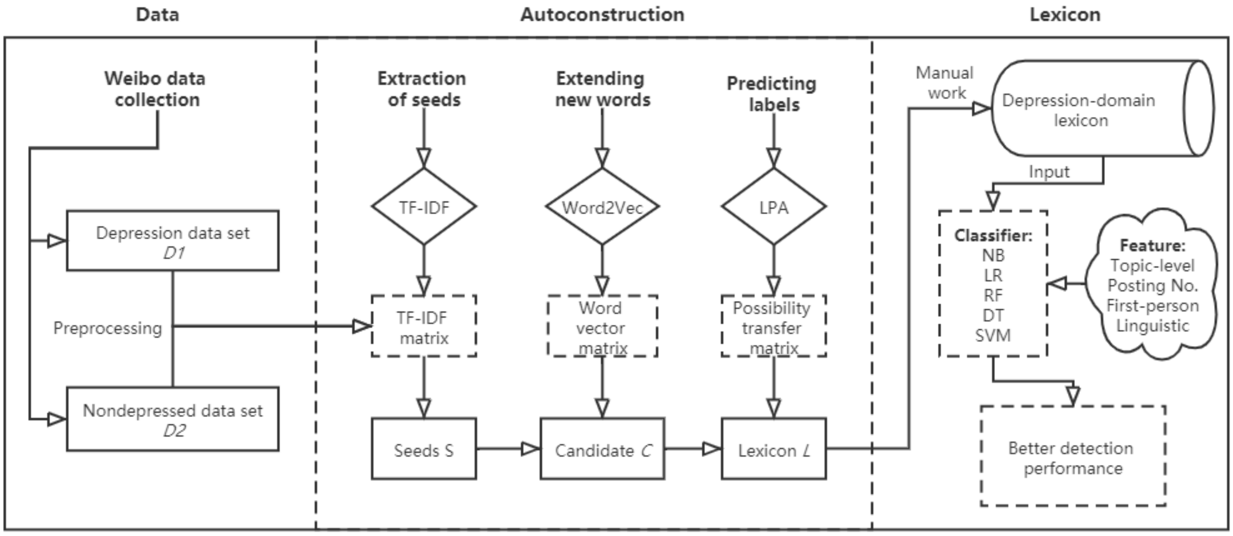
Construction Overview

Domain adaptability is always a difficult problem in natural language processing. Therefore, a domain-based lexicon can help us perform analysis in a more accurate and deeper way. For example, "excitement," "life," and "forever" are common words in our daily life, but they can be abnormal signals of a patient with depression. Thus, through our study, we will try our best to determine which words used on the internet indicate depression and which do not indicate depression.

There are many ways to construct a domain-based lexicon according to a survey [38], which can mainly be divided into knowledge base, corpus base, and these two combined. Knowledge base includes traditional methods, such as word relation extension [39] and annotation extension [40]. Corpus base refers to conjunction syntax [41] and word co-occurrence [42]. In fact, a survey [38] showed that the class of methods adopted more widely is the automatic method combining existing knowledge and corpus base. In this view, approaches involving semisupervised construction on relationship graphs like the LPA, bootstrapping, and word embedding are popular and effective methods mentioned in the subsection Related Work About Research on Construction of a Domain Lexicon.

Inspired by Hamilton [20], Giulianelli [33], and Pu et al [34], we applied automatic construction, a semisupervised method that combines W2V and LPA on a lexical semantic relationship graph, to obtain a depression-domain lexicon containing depressed feature words. Our construction can be divided into the following four steps: (1) *Extraction of seed words*. Extract a few words that are most important and valuable in the domain; (2) *Extending new words*. Use the W2V model to learn word vectors on the corpus and then extend with similarity between seeds and new words; (3) *Setting labels for the new words*. If the cosine similarity of a word and a seed is greater than the threshold, an edge will be formed, and the weight of the edge is the similarity. Through such iteration, a graph is obtained. After that, the LPA is used on the semantic graph to obtain the labels of all candidate words; (4) *Obtaining the domain-based lexicon*. Finally, by simple manual arrangement, the depression-domain lexicon is formed. We then used it as an input for detection models and found that the models performed much better than before. The method is described as a detailed framework in Figure 1.

Figure 1. An illustration of the framework. DT: decision tree; LR: logistic regression; NB: naive Bayes; RF: random forest; SVM: support vector machine; TF-IDF: term frequency-inverse documentation frequency.



Extraction of Seed Words

Seed words are those that can be representative of a specific domain. In order to extract the key seed words in the depressed and nondepressed data sets, we leveraged the TF-IDF algorithm, which is a widely used feature extraction algorithm in natural language processing. Salton and Yu [43] first proposed the TF-IDF algorithm, and Salton et al [44] demonstrated its validity in information retrieval. Term frequency (TF) refers to the number of times a term or word occurs in a document, and inverse document frequency (IDF) is related to the frequency of a term appearing in all documents, which measures specificity of the term over the entire corpus.

TF and IDF can be formulated as follows:

where $n_{i,j}$ is the word i in document j, k is the number of words in j, N is the number of documents containing word i, D is the size of the documents, and $DF(i)$ is the number of documents in which the word i occurs at least once. Additionally, *tfidf* can be formulated as follows:

Intuitively, this calculation of TF-IDF will show us how important and special a given word is in our depression domain. Words with a higher *tfidf* value tend to have a greater relevance in a document. In our research, we regarded the data sets $D1$ and $D2$ as two corpora and every microblog as a document. We then extracted words with the highest TF-IDF values in our corpora.

Extending New Words With W2V

Now that we had the seeds S, we could leverage the word embedding model to extend new words. Word embeddings, which help map the vocabulary to vectors, are popular tools for natural language processing. We adopted W2V, an efficient algorithm for learning embeddings using a neural language model, to generate the vectors. W2V is an open source model by Mikolov et al [45] at Google, and its main idea is to use deep learning technology on a specific corpus and then to map each word into a multidimensional real vector space, where the distance between words that have a higher semantic similarity is small.

In this paper, cosine similarity was used to calculate the similarity between words. When a word whose similarity with the seed words in the training corpus was greater than the given threshold, we extracted it as a new word and added it as a candidate word to the candidate word set C. If S_i and C_j represent the vectors of a seed word and candidate word, respectively, the similarity between them can be formulated by $SIM(Si, Cj)$ as follows:

Setting Labels With Label Propagation

The LPA is a common semisupervised approach on a graph [28]. It has been applied to many fields, such as community detection [46], personal social relation extraction [47], and dictionary construction. Using a graph model to construct a lexicon can capture the global relations among all words, overcome the dependence on seeds, and provide a better result in the case of limited labeled data.

The LPA builds a graph based on the similarity between nodes, which are the words in our study. After the initialization of the graph, the nodes in the graph can be divided into labeled nodes and unknown nodes. The basic idea of LPA is to predict the label of unknown nodes based on information from labeled ones, and labels are propagated mainly by the weight on the edge between the nodes. In the process of label propagation, unknown nodes can update their own labels through the information of adjacent known labels. If the similarity of the adjacent node is large, the influence of its label will be large.

In our algorithm, the seeds *S* are taken as the labeled nodes and the extended candidate words *C* are taken as the unknown nodes. The semantic graph is constructed as follows: If the seed word *i* is extended by W2V to get a new word *j*, there is an edge between *i* and *j*, and the weight of the edge is the similarity of the two words. Thus, all of the seed words and candidate words will form a semantic graph as shown in Figure 2.

Assuming that there are *n* nodes in total, then an *n*-dimensional transition probability matrix can be constructed. Let $SIM(w_i,w_j)$ represent the similarity between w_i and w_j, which is calculated by cosine similarity. $T[i][j]$represents the similarity transfer probability from word *i* to word *j*, which is calculated as follows:

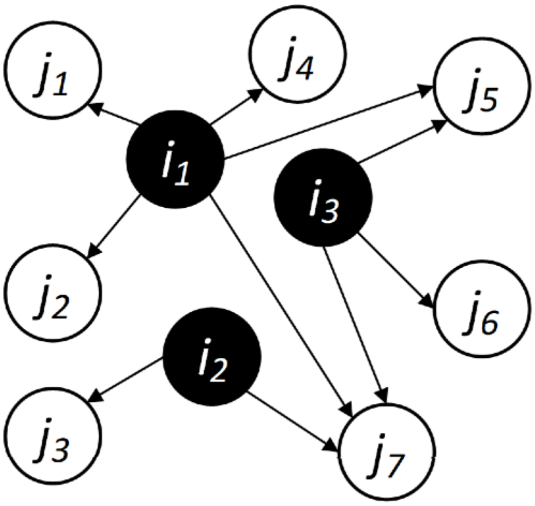
If there are 10 nodes in the graph, in which i_1 and i_2 are depression seed words with the label "−1," $i_3$is a nondepression seed word with the label "+1," and the labels of the rest of the candidate words are unknown (given an initial value of 0), the initial labels of all nodes can be represented by the vector *V* as follows:

The label of each unknown candidate word is obtained by iteratively applying the transition matrix on the initial labels of the words. The calculation method is as follows:

where *Label* represents the label possibility of node *j* after the iteration, $T[i][j]$ represents the transfer probability in the similarity matrix of node *i* to node *j*, and $V[i]$ represents the initial *Label* of node *i* before the iteration.

In each iteration, the labels of the seeds should remain the same. When the labels of all words in the graph no longer change after continuous iteration, the iteration is over. At the end of the iteration process, the final candidate words are those words whose absolute value of label probability is greater than a certain threshold. In this way, we obtained a well-labeled domain lexicon. The previous algorithms can be concluded as the steps in Textbox 1.

Figure 2. A simple structure of a semantic graph. i: seed word; j: candidate word.



Textbox 1. Algorithms of the procedure.

Input: Corpus of data set (Corpus=$D1 \cup D2$), seeds S, and the threshold T_c for candidate words C.

Output: One depression-domain lexicon L with depressive words L_d and nondepressive words L_n.

Procedure:

1) Initialize the lexicon and candidate words. $C=\varnothing$, $L_d=\varnothing$, $L_n=\varnothing$.

2) Preprocess the corpus and learn the word embedding with Word2Vec.

3) For every seed, $S_i \in S$:

For a word W_j in *Corpus*, if $SIM(S_i,W_j) \geq T_c$, then $C=C \cup S_i \cup W_j$. Record the similarity calculated by equation (4).

4) After obtaining all the extended candidate words C and the similarity matrix between words through step 3), the transition probability matrix of similarity in C can be constructed according to equation (5). Then, the semantic relationship graph is obtained.

5) In the whole graph, *Label* of unknown words is calculated according to formula (7) given the initial label V.

6) Reset the labels of the seeds in *Label* to its initial value. Then, let V=*Label*.

7) Repeat steps 5) and 6) until the labels of C in the graph do not change anymore.

8) Obtain the final *Label*. For $C_i \in C$, if $Label_{Ci}$ <0 and $|Label_{Ci}|$ >0.5, then $L_d=L_d \cup C_i$; For $C_j \in C$, if $Label_{Cj}$ >0 and $|Label_{Cj}|$ >0.5, then $L_n=L_n \cup C_j$.

9) Combine L_d and L_n and finally obtain the depression-domain lexicon L after manual work.

Results

Experiment Setup

We employed our data set to construct a depression-domain lexicon. We needed two types of microblogs combining users with depression or those without depression to extract domain seed words and to finish the automatic construction with the LPA. Our original data crawled from Weibo had some noise, especially in *D*1, so manual preprocessing (detailed description in the “Data Preprocessing” subsection) was necessary to clean the data into *D*1 and *D*2.

After our lexicon was automatically built, we labeled it depressed or nondepressed for further evaluation. Three volunteers, who had carefully read the depressed microblogs and research articles, were invited to perform the lexicon labeling job [48]. Thus, every word in the lexicon was labeled three times. If there was a labeling disagreement, voting was adopted to obtain the ground truth.

Word Segmentation

Chinese word segmentation has a great influence in lexicon construction, especially when it comes to Weibo microblogs and the depression domain. In order to segment Chinese words properly in Weibo text, we used the following three steps to segment words as accurately as possible: (1) domain dictionary; (2) large word embedding; and (3) incorrect word removal.

Domain Dictionary

When coping with mental disease, especially depression, over the internet, some depression-domain words like paroxetine (“帕罗西汀”), which is a common antidepressant, and self-rating scale (“自评量表”), which is a tool for individuals to measure depression, were difficult to recognize. Other words like MLGB (“马勒戈壁”), which means damn it, and Yali (“鸭梨”), which means pressure, were network vocabularies that could be confusing for the computer. Domain-specific segmentation should combine a domain dictionary [49]; however, there is no depression dictionary in public resources. To solve the segmentation problem, we downloaded “Dictionary of Psychology” and “Dictionary of Neuropsychiatry” from the CNKI Tool library [50] (there is no depression lexicon yet, so we chose the dictionary of psychology and psychiatry; CNKI is one of the largest Chinese knowledge discovery web-based platforms), downloaded “Weibo Dictionary” from BosonNLP [51] (a dictionary automatically constructed from millions of annotation data points from microblogs, forums, and other data sources), and used a manually collected antidepressant dictionary [26] (words like amitriptyline and paroxetine in our data sets were replaced with antidepressant as a data reduction method) from web-based pharmacies and science articles. The work of Chinese domain word segmentation was inspired by Fang [26] and Cheng [49]. The final domain dictionary contained 122,594 words after eliminating duplicate words. We then used jieba (built to be the best Python Chinese word segmentation module) [52] as our segmentation module, which adopted the unigram model and hidden Markov model.

Large Word Embedding

A richer corpus is associated with more precise word embedding. Instead of using our collected data, which were relatively rare, we leveraged the W2V models by Shen et al [53], which are trained on 5 million Weibo microblogs and 223 million Chinese Wiki tokens, for word embeddings.

Incorrect Word Removal

We planned to remove incorrect words from our lexicon. Actually, after evaluation, we found that the error rate was less than 2% to 3%. Among 2385 words in our depression-domain lexicon, there were 64 errors.

Evaluation Metrics

During our experiments, we constructed the depression-domain lexicon with an automatic method, compared our method with some baseline approaches, and analyzed key parameters like number of seeds and threshold in the model.

For the evaluation metrics, we employed precision, recall, and F1 measure (F1) in equations (8), (9) and (10), respectively, to evaluate the performance of our model and the baseline approaches. We used area under the curve (AUC) to evaluate the model of unbalanced data. In terms of the number of words in the lexicon, we also compared the numbers under different circumstances. The equations are as follows:

where *TP* represents true positive, which means depressed words are correctly detected as depressed; *FN* is false negative, which means depressed words are incorrectly determined as nondepressed; and *FP* is false positive, which means that nondepressed words are incorrectly detected as depressed.

Figure 1 provides an entire picture of the experiment.

Seed Words

Before construction, we used the TF-IDF to extract the seed words and obtained a list of the top 2000 words. The samples of the TF-IDF of *D*1 are shown in Table 2.

By artificially screening the list, we could obtain some seed words. Moreover, we added a few general sentiment words with high levels to our seed words and finally obtained a set of seed words of 40 depressive seeds and 40 nondepressive seeds. From parameter sensitivity analysis, we noted that 80 seeds in total will lead to a sufficiently large lexicon with high accuracy. The samples of the 80 seeds are shown in Table 3.

Table 2. TF-IDF values of depressed D1 samples.

Depressed *D*1	TF-IDF[a] value
Myself (自己)	0.041383
Really (真的)	0.032475
Depression (抑郁症)	0.024328
Hope (希望)	0.013336
Life (生活)	0.012043
Forever (永远)	0.006965
Pain (痛苦)	0.006871
Sad (难过)	0.006756
Live (活着)	0.006583
Mood (心情)	0.006386
Night (晚上)	0.006347
Always (总是)	0.005984
Hate (讨厌)	0.005475
Exhausted (好累)	0.005469
Fear (害怕)	0.005030
Lonely (孤独)	0.004413
Idiot (傻逼)	0.004380
Emotion (感情)	0.004031
Insomnia (失眠)	0.003950
Sorry (对不起)	0.003867
Despair (绝望)	0.003410
Antidepressant (抗抑郁药)	0.002305

[a] TF-IDF: term frequency-inverse documentation frequency.

Table 3. Summary of the seeds.

Category	Seeds *S*
Nondepressive (40 words)	Stability, comfort, happy, happiness, successful, confidence, sunshine, struggle, positive, brave, enjoy, peace, enthusiasm, healthy, satisfied, active, grow up, pride, good, admire, strong, perfect, praise, precious, progress, congratulate, love, welcome, kindness, robust, earnest, agree, support, award, advantage, good deal, develop, warm, bright colored, and understand (稳定, 舒服, 高兴, 幸福, 顺利, 自信, 阳光, 奋斗, 积极, 勇敢, 享受, 平安, 热情, 健康, 满意, 活力, 成长, 骄傲, 优秀, 敬佩, 完美, 称赞, 强大, 珍贵, 进步, 庆贺, 关爱, 欢迎, 强壮, 善良, 认真, 同意, 支持, 奖励, 优势, 划算, 发展, 温暖, 鲜艳, 明白)
Depressive (40 words)	Depression, collapse, stress, suicide, apastia, anxious, sad, tired, death, lonely, insomnia, bad, desperate, give up, low, leave, fear, danger, close, sensitive, lost, shadow, destroy, suspect, crash, dark, helpless, guilt, negative, frustration, nervous, melancholy, rubbish, jump, forget, goodbye, cut wrist, edge, haze, and antidepressant (抑郁, 崩溃, 压力, 自杀, 绝食, 焦躁, 伤心, 疲惫, 死亡, 孤独, 失眠, 难受, 绝望, 放弃, 卑微, 离开, 恐惧, 危险, 封闭, 敏感, 茫然, 阴影, 摧毁, 怀疑, 崩塌, 黑暗, 无助, 愧疚, 负面, 沮丧, 紧张, 忧郁, 废物, 跳楼, 遗忘, 再见, 割腕, 边缘, 阴霾, 抗抑郁药)

Model Evaluation

In order to verify the effectiveness of the lexicon autoconstruction method applied in this paper, we selected the following methods as baseline approaches: (1) *W2V* [24-26]. A common method of constructing a lexicon based on W2V, which is used to learn word embedding vectors on a corpus. The semantic similarity between words and seed words in the corpus is then iteratively calculated. If the similarity is greater than a certain threshold, the new word is extended and has the same label as the seed word; (2) *SO-W2V* [27]. It is a semisupervised sentiment orientation classification algorithm based on a word vector. The basic idea is that through comparison with all positive and negative seed words, an accurate orientation of the extended word will be obtained. It has versatility in different areas for a Chinese corpus; (3) *SO-PMI* [9,22,23]. It calculates the probability of the occurrence of both seed words and expanded words in the text. A higher probability is associated with a closer correlation; (4) *W2V-LPA*. It is our method, which considers both the word relationship and specific domain context.

To obtain a fair comparison, we set the same parameters for all methods where T_c was 0.5 and the size of seeds S was 80. For W2V tools, we used the gensim package [54].

From Table 4 and Figure 3, we can see the evaluation results. It is obvious that the W2V-LPA and W2V methods performed much better than the SO-W2V and SO-PMI methods. Moreover, when the size of seeds increased from 60 to 120, our method was able to maintain a more stable and precise performance, which was almost 1% to 6% higher than others (Figure 3), whereas the value for SO-W2V declined quickly when the size of seeds became larger. Overall, SO-W2V takes all the other seeds into account, but too many seeds combined will introduce too much noise to some extent, as not all seeds are related to an extended word. W2V is a simple and general method, which only considers the label of the first seed when extending new words. Additionally, SO-PMI mainly takes word co-occurrence frequency into account. What W2V-LPA did better is that it only predicted labels through the semantic graph of related and similar words, and thus, the semantic context and word relation were both considered. Therefore, we can say that W2V-LPA is a much better and more stable method for the autoconstruction of a domain lexicon.

Table 4. Performance of lexicon construction methods.

Construction method	Precision	Recall	F1	Size of the lexicon
W2V-LPA[a]	0.880	0.906	0.893	2321
W2V[b]	0.878	0.903	0.890	2321
SO-PMI[c]	0.879	0.877	0.877	2024
SO-W2V[d]	0.854	0.877	0.862	2321

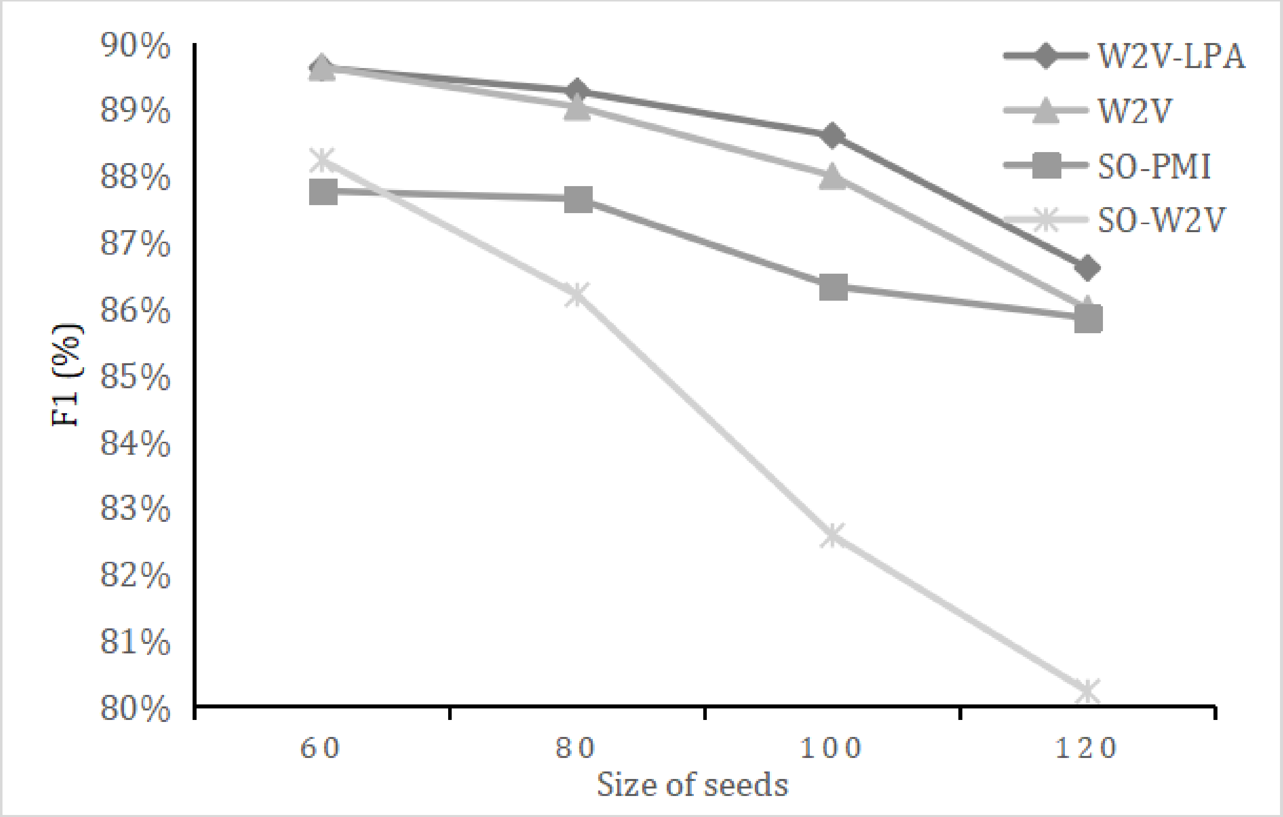
[a] W2V-LPA: label propagation algorithm-Word2Vec.

[b] W2V: Word2Vec.

[c] SO-PMI: semantic orientation pointwise mutual information.

[d] SO-W2V: semantic orientation Word2Vec.

Figure 3. F1 of methods when the seed size changed. LPA: label propagation algorithm; SO-PMI: semantic orientation pointwise mutual information; SO-W2V: semantic orientation from Word2Vec; W2V: Word2Vec.



Parameter Sensitivity Analysis

Throughout our experiment, the size of seeds *S* and the extension threshold T_c were two important parameters. More seeds or a lower threshold will lead to a lexicon with more words but lower accuracy, whereas fewer seeds and a high threshold will ensure more precision but a poor lexicon. We balanced the trade-offs, as we wanted to obtain a relatively accurate and abundant lexicon that would be helpful for further depression diagnosis. Figure 4 presents the size of the lexicon when the size of seeds and threshold for candidate words changed.

First, we fixed T_c at 0.7 and then varied the size of seeds from 60 to 120. If we have less than 60 seeds, the entire lexicon will be so small that almost nothing will remain but seed words. A size larger than 60 will not change the outcome, so 0.7 might be a very high-level threshold. From Table 5, we can see that larger sizes of seeds like 100 and 120 partially jeopardized the performance, and W2V-LPA performed nearly the same when the sizes were 60 and 80.

We then fixed the size of seeds at 80 with varying T_c from 0.7 to 0.5. With a higher threshold, the performance was relatively excellent, whereas the size of the lexicon started to fail at around 1000 when T_c was 0.55. We believe a lexicon with 2000 words and a T_c of 0.5 might have good balance.

Overall, it is pleasing that our W2V-LPA method performed quite smoothly and steadily even when the parameters were changed, so we believe that a high-quality lexicon can be constructed. It is difficult to find an optimal solution, and given *D*1 and *D*2, we will adopt a size of seeds of 80 and a threshold T_c of 0.5 as a relatively proper approach.

Figure 4. Size of the lexicon when the size of seeds and threshold for candidate words changed.

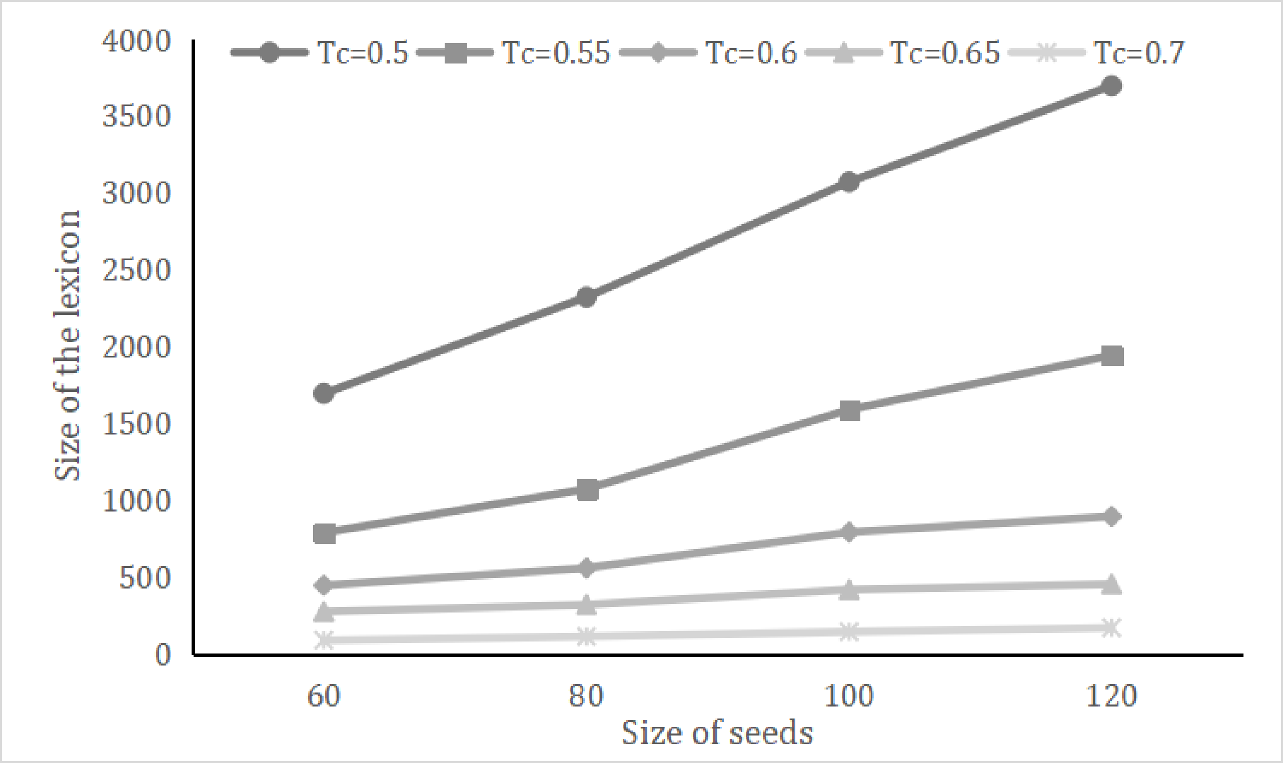


Table 5. Performance of the W2V-LPA method when S and T_c were changed.

S [a]	T_c [b]	Precision	Recall	F1	Size of the lexicon
60	0.5	0.882	0.911	0.896	1694
60	0.55	0.910	0.935	0.922	788
60	0.6	0.926	0.944	0.935	446
60	0.65	0.951	0.963	0.954	275
60	0.7	0.804	0.897	0.848	89
80	0.5	0.880	0.906	0.893	2321
80	0.55	0.916	0.937	0.926	1072
80	0.6	0.934	0.948	0.941	558
80	0.65	0.954	0.963	0.958	320
80	0.7	0.918	0.909	0.892	113
100	0.5	0.874	0.899	0.886	3070
100	0.55	0.906	0.924	0.915	1589
100	0.6	0.927	0.937	0.931	792
100	0.65	0.953	0.959	0.955	418
100	0.7	0.937	0.932	0.925	144
120	0.5	0.855	0.879	0.866	3696
120	0.55	0.889	0.904	0.896	1942
120	0.6	0.924	0.933	0.928	894
120	0.65	0.952	0.958	0.954	454
120	0.7	0.944	0.940	0.934	170

[a] S: size of the seeds.

[b] T_c: threshold for candidate words.

Detection Performance

After construction of the depression-domain lexicon, we could apply it to actual depression detection in a new Weibo microblog data set to find out if our work would help existing detection models perform better. The detection process included data collection, feature selection, and classification methods.

Data Collection

In addition to our data set used for lexicon construction, we collected 745 users who were depressed and 10,118 users who were not depressed with their 1-year tweets as a new data set. Data details are shown in Table 6.

Table 6. Details of the data set for depression detection.

Data set	Users	Total posts	Mean	Standard deviation	Skewness	Kurtosis	Time span
Depressed data set *D*3	745	179,600	240.44	486.28	6.21	56.32	January 2018-June 2019 (18 months)
Nondepressed data set *D*4	10,118	3,150,000	310.93	327.72	3.50	48.52	January 2018-June 2019 (18 months)

Feature Selection

Features like topic-level keywords, posting behaviors, number of tweets, first-person words, and linguistic style are meaningful in detecting depression on the internet [11,13]. We also set our depression-domain lexicon as one feature to see whether it would really contribute a lot after inclusion in the detection model. The features were as follows: (1) *Topic-level keywords*. We selected 30 topic-level keywords with the TF-IDF; (2) *Posting behaviors*. For each user, average length of tweets and total posting numbers were collected to represent web-related posting behaviors; (3) *First-person words*. According to linguistic inquiry and word count [55], we counted the number of first-person pronouns like I, we, us, etc; (4) *Linguistic style (200 dimensions)*. To approximately analyze linguistic style, we calculated the average vectors of every user with Word2Vec [56]. Finally, we constructed the depression-domain lexicon by the previously mentioned process.

Classification Methods

We chose naive Bayes (NB), decision tree, logistic regression (LR), random forest, and support vector machine [5,37] as classification methods to detect users with depression. From model performance, we obtained a quick picture about the importance of our lexicon. When the depression-domain lexicon is selected as one feature, the method has the tag L. For example, L-NB is a classification algorithm that has the feature of the depression-domain lexicon, whereas NB does not have this feature. After including the depression-domain lexicon in the models, we clearly found that each detection performance improved when compared with before inclusion of the lexicon (Table 7). The performance of lexicon methods surpassed that of corresponding methods without the lexicon by 2% to 9%, which justifies the important role of our lexicon in depression detection.

The model was based on a data set with 50% users who were depressed and 50% users who were not depressed. When we varied the scale of depressed users, the data set became imbalanced and the AUC was more important to test the performance. Figure 5 illustrates the trend of detecting performance when setting different proportions of users who were depressed in the L-LR method. This method achieved an outstanding performance when the proportion of users with depression was 50%. However, the AUC dropped sharply when the data set was imbalanced.

In the real word, people with depression make up less than 10% of the population, and we will determine how to properly detect depression with imbalanced data in a further study.

Table 7. Detection model performance with the depression-domain lexicon.

Detection model	Precision	Recall	F1	Accuracy
NB[a]	67%	67%	67%	67%
L[b]-NB	74%	73%	73%	73%
LR[c]	76%	76%	75%	76%
L-LR	77%	77%	77%	77%
RF[d]	68%	68%	68%	68%
L-RF	77%	77%	76%	77%
SVM[e]	65%	65%	65%	65%
L-SVM	74%	72%	72%	72%
DT[f]	67%	67%	67%	67%
L-DT	69%	69%	69%	69%

[a]NB: naive Bayes.

[b]L: depression-domain lexicon as a feature.

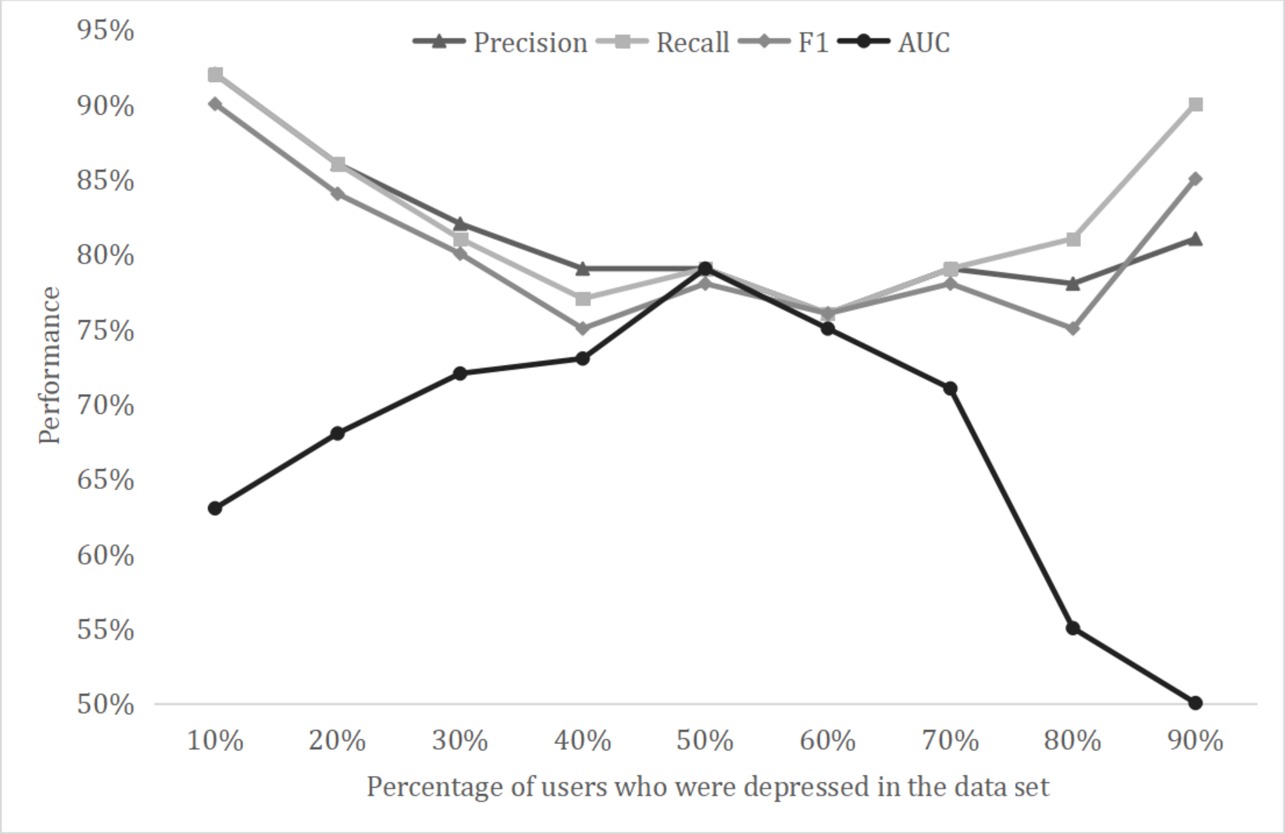
[c]LR: logistic regression.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]DT: decision tree.

Figure 5. Scales of users who were depressed. AUC: area under the curve.



Discussion

Diagnosis of users with potential depression via social media has attracted increasing attention because it is a more cost-effective and active approach dealing with massive valuable data than traditional diagnosis. In previous studies, most of the achievements about a lexicon involved an English corpus. Instead of translating an English lexicon, this paper aimed to apply an automatic construction method for a Chinese depression-domain lexicon based on the LPA. With Word2Vec and a semantic relationship graph, the LPA was used to predict the label of candidate words in the graph, and finally, our lexicon was constructed. Experiment results showed that our method was superior to baseline construction methods and had good performance and robustness. In addition, when our lexicon was included as an input for the detection models, their performance became more accurate and effective when compared with the models without the depression-domain lexicon.

In the next step, experiments are expected to be carried out on a larger depression corpus, and more linguistic knowledge like conjunction will be incorporated into our method to enlarge the range of the depression-domain lexicon. Meanwhile, more complex construction methods like deep neural networks and hierarchical topic models will be adopted in further research. We expect that our lexicon will act as a useful feature in depression detection and will be able to provide more insights for depression diagnosis in terms of advanced depression detection among patients.

Acknowledgments

GL led the method application, experiment conduction, and result analysis. LH and SH participated in data extraction, preprocessing, and manuscript revision. BL provided theoretical guidance and revised the paper. This work was supported by the National Social Science Fund Project, China (No. 16BTQ065) “Multi-source intelligence fusion research on emergencies in big data environment” and the Foundation for Disciplinary Development of the School of Information and Technology in the University of International Business and Economics.

Conflicts of Interest

None declared.

References

1. World Health Organization. World Health Statistics 2017: Monitoring health for the SDGs URL: https://www.who.int/gho/publications/world_health_statistics/2017/en/ [accessed 2020-06-05]
2. Qin X, Wang S, Hsieh C. The prevalence of depression and depressive symptoms among adults in China: Estimation based on a National Household Survey. China Economic Review 2018 Oct;51:271-282. [doi: 10.1016/j.chieco.2016.04.001]
3. Lépine JP, Briley M. The increasing burden of depression. NDT 2011 May;7:3-7. [doi: 10.2147/ndt.s19617]
4. O'Loughlin K, Neary M, Adkins EC, Schueller SM. Reviewing the data security and privacy policies of mobile apps for depression. Internet Interv 2019 Mar;15:110-115 [FREE Full text] [doi: 10.1016/j.invent.2018.12.001] [Medline: 30792962]
5. Rahman RA, Omar K, Noah SA, Mohd Shahrul Nizam Mohd Danuri MS. A Survey on Mental Health Detection in Online Social Network. International Journal on Advanced Science, Engineering and Information Technology 2018;8:1431-1436. [doi: 10.18517/ijaseit.8.4-2.6830]
6. Lin H, Jia J, Qiu J, Zhang Y, Shen G, Xie L, et al. Detecting Stress Based on Social Interactions in Social Networks. IEEE Trans. Knowl. Data Eng 2017 Sep 1;29(9):1820-1833. [doi: 10.1109/tkde.2017.2686382]
7. Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A. Recognition of depression by non-psychiatric physicians--a systematic literature review and meta-analysis. J Gen Intern Med 2008 Jan;23(1):25-36 [FREE Full text] [doi: 10.1007/s11606-007-0428-5] [Medline: 17968628]
8. Park MS, Chiyoung C, Meeyoung C. Depressive moods of users portrayed in twitter. 2012 Presented at: Proceedings of the ACM SIGKDD Workshop On Healthcare Informatics (HI-KDD); 2012; San Diego p. 1-8.
9. Choudhury D, Gamon M, Counts M, Horvitz S. Predicting Depression via Social Media. 2013 Jul Presented at: International AAAI Conference on Weblogs and Social Media; 2013; Cambridge, Massachusetts, USA p. 128-137.
10. Hasan M, Rundensteiner E, Agu E. EMOTEX: Detecting Emotions in Twitter Messages. 2014 Apr Presented at: ASE BigData/SocialCom/CyberSecurity Conference; 2014; Stanford, California, USA.
11. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, Boyd-Graber J. Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. 2015 Presented at: The 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; 2015; Denver, Colorado, USA p. 99-107. [doi: 10.3115/v1/w15-1212]
12. Tsugawa S, Kikuchi S, Kishino Y, Nakajima F, Itoh K, Ohsaki H. Recognizing Depression from Twitter Activity. In: ACM. 2015 Presented at: The 33rd Annual ACM Conference on Human Factors in Computing Systems; 2015; Seoul, Republic of Korea p. 3187-3196. [doi: 10.1145/2702123.2702280]
13. Guangyao S, Jia J, Liqiang N, Fuli F, Cunjun Z, Tianrui H, et al. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. 2017 Presented at: The Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017; Melbourne, Australia p. 3838-3844. [doi: 10.24963/ijcai.2017/536]
14. Losada DE, Gamallo P. Evaluating and improving lexical resources for detecting signs of depression in text. Lang Resources & Evaluation 2018 Aug 6;54(1):1-24. [doi: 10.1007/s10579-018-9423-1]
15. GitHub. Chinese-Depression-domain-Lexicon URL: https://github.com/omfoggynight/Chinese-Depression-domain-Lexicon [accessed 2020-04-15]

16. Beck AT. An inventory for measuring depression. Arch Gen Psychiatry 1961 Jun;4:561-571. [doi: 10.1001/archpsyc.1961.01710120031004] [Medline: 13688369]
17. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, Texas: Psychological Corporation; 1996.
18. Radloff LS. The CES-D Scale. Applied Psychological Measurement 2016 Jul 26;1(3):385-401. [doi: 10.1177/014662167700100306]
19. Zung WW. A self-rating depression scale. Arch Gen Psychiatry 1965 Jan;12:63-70. [doi: 10.1001/archpsyc.1965.01720310065008] [Medline: 14221692]
20. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967 Dec;6(4):278-296. [doi: 10.1111/j.2044-8260.1967.tb00530.x] [Medline: 6080235]
21. Hu RJ. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). Encyclopedia of the Neurological Sciences 2003;25(2):4-8.
22. Das D, Poria S, Bandyopadhyay S. A Classifier Based Approach to Emotion Lexicon Construction. Natural Language Processing and Information Systems 2012;7337. [doi: 10.1007/978-3-642-31178-9_41]
23. Krestel R, Siersdorfer S. Generating Contextualized Sentiment Lexica based on Latent Topics and User Ratings. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. 2013 Presented at: The 24th ACM Conference on Hypertext and Social Media; 2013; Paris, France p. 01-2013. [doi: 10.1145/2481492.2481506]
24. Yu M, Dredze M. Improving Lexical Embeddings with Semantic Knowledge. 2014 Presented at: The 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); 2014; Baltimore, Maryland, USA. [doi: 10.3115/v1/p14-2089]
25. Tixier AJ, Vazirgiannis M, Hallowell MR. Word Embeddings for the Construction Domain. arXiv e-prints 2016 [FREE Full text]
26. Zhenyu F. Research on depression prediction of micro-blog users based on word embedding method. Electronic Technology & Software Engineering 2017.
27. Chao F, Xun L, Yaping L. Construction Method of Chinese Cross-Domain Sentiment Lexicon Based on Word Vector. Journal of Data Acquisition and Processing 2017:579-587.
28. Xiaojin Z, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation. Tech Report 2002.
29. Rao D, Ravichandran D. Semi-Supervised Polarity Lexicon Induction. 2009 Presented at: 12th Conference of the European Chapter of the Association for Computational Linguistics; 2009; Athens, Greece. [doi: 10.3115/1609067.1609142]
30. Brody S, Elhadad N. An Unsupervised Aspect-Sentiment Model for Online Reviews. 2010 Presented at: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics; 2010; Los Angeles, California, USA.
31. Yen-Jen T, Hung-Yu K. Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation. In: Proceedings of International Conference on Information Integration and Web-based Applications & Services. 2013 Presented at: Conference on Information Integration and Web-based Applications & Services; 2013; Vienna, Austria p. 02-2013. [doi: 10.1145/2539150.2539190]
32. Hamilton WL, Clark K, Leskovec J, Jurafsky D. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. Proc Conf Empir Methods Nat Lang Process 2016 Nov;2016:595-605. [doi: 10.18653/v1/D16-1057] [Medline: 28660257]
33. Giulianelli M. Semi-supervised emotion lexicon expansion with label propagation and specialized word embeddings. arXiv e-prints 2017 [FREE Full text]
34. Pu Z, Junxia W, Yinghao W. Sentiment Lexicon Construction Method Based on Label Propagation. Computer Engineering 2018;44(5):168-173.
35. Sina Weibo. 2018 Weibo User Development Report URL: https://data.weibo.com/report/reportDetail?id=433 [accessed 2020-06-05]
36. Coppersmith G, Mark D, Craig H. Quantifying Mental Health Signals in Twitter. 2014 Presented at: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; 2014; Baltimore, Maryland, USA p. 51-60. [doi: 10.3115/v1/W14-3207]
37. Tiancheng S, Jia J, Guangyao S, Fuli F, Xiangnan H, Huanbo L, et al. Cross-Domain Depression Detection via Harvesting Social Media. 2018 Presented at: Twenty-Seventh International Joint Conference on Artificial Intelligence; 2018; Stockholm p. 1611-1617. [doi: 10.24963/ijcai.2018/223]
38. Ke W, Rui X. A survey on automatical construction methods of sentiment lexicons. Acta Automatica Sinica 2016;42(4):495-511. [doi: 10.16383/j.aas.2016.c150585]
39. Blair-goldensohn S, Neylon T, Hannan K, Reis RA, Mcdonald R, Reynar J. Building a Sentiment Summarizer for Local Service Reviews. 2008 Presented at: WWW2008 Workshop: NLP in the Information Explosion Era (NLPIX 2008); 2008; Beijing, China.
40. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. 2010 Presented at: International Conference on Language Resources and Evaluation; 2010; Valletta, Malta p. 17-23.

41. Kanayama H, Nasukawa T. Fully Automatic Lexicon Expansion for DomainOriented Sentiment Analysis. 2006 Presented at: Conference on Empirical Methods in Natural Language Processing; 2006; Sydney, Australia. [doi: 10.3115/1610075.1610125]
42. Krestel R, Siersdorfer S. Generating contextualized sentiment lexica based on latent topics and user ratings. 2013 Presented at: The 24th ACM Conference on Hypertext and Social Media; 2013; Paris, France p. 129-138. [doi: 10.1145/2481492.2481506]
43. Salton G, Yu CT. On the construction of effective vocabularies for information retrieval. SIGPLAN Not 1975 Jan 01;10(1):48-60. [doi: 10.1145/951787.951766]
44. Salton G, Fox EA, Wu H. Extended Boolean information retrieval. Commun. ACM 1983;26(11):1022-1036. [doi: 10.1145/182.358466]
45. Mikolov T, Sutskever I, Chen K. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 2013.
46. Cordasco G, Gargano L. Community Detection via Semi-Synchronous Label Propagation Algorithms. Int. J. of Social Network Mining 2012. [doi: 10.1109/basna.2010.5730298]
47. Boldi P, Rosa M, Santini M, Vigna S. Layered label propagation: a multiresolution coordinate-free ordering for compressing social networks. 2011 Presented at: The 20th International Conference on World Wide Web; April 01, 2011; Hyderabad, India. [doi: 10.1145/1963405.1963488]
48. Huijie L, Jia J, Liqiang N, Guangyao S, Tat-Seng C. What does social media say about your stress. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016 Presented at: The Twenty-Fifth International Joint Conference on Artificial Intelligence; 2016; New York, USA p. 3775-3781.
49. Yusi C, Yuntao S. Domain specific Chinese word segmentation. Computer Engineering and Applications 2018;54(17):30-34.
50. CNKI. CNKI tool library URL: http://mall.cnki.net/reference/index.aspx [accessed 2020-04-15]
51. BOSON. BosonNLP Dictionary URL: https://bosonnlp.com/dev/resource [accessed 2018-09-01]
52. GitHub. jieba URL: https://github.com/fxsjy/jieba [accessed 2018-09-01]
53. Shen L, Zhe Z, Renfen H, Wensi L, Tao L, Xiaoyong D. Analogical Reasoning on Chinese Morphological and Semantic Relations. 2018 Presented at: The 56th Annual Meeting of the Association for Computational Linguistics; 2018; Melbourne, Australia.
54. GitHub. Gensim – Topic Modelling in Python URL: https://github.com/RaRe-Technologies/gensim [accessed 2018-09-01]
55. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count (LIWC). Mahwah, NJ: Erlbaum Publishers; 1999.
56. Dinkel H, Wu M, Yu K. Text-based Depression Detection: What Triggers An Alert. arXiv e-prints 2019 [FREE Full text]

Abbreviations

DT: decision tree
LPA: label propagation algorithm
LR: logistic regression
NB: naive Bayes
RF: random forest
SO-PMI: semantic orientation pointwise mutual information
SO-W2V: semantic orientation from Word2Vec
SVM: support vector machine
TF-IDF: term frequency-inverse documentation frequency
W2V: Word2Vec

Edited by T Hao, B Tang; submitted 31.12.19; peer-reviewed by Z Zhang, D Carvalho; comments to author 20.02.20; revised version received 11.04.20; accepted 05.05.20; published 23.06.20.

Please cite as:
Li G, Li B, Huang L, Hou S
Automatic Construction of a Depression-Domain Lexicon Based on Microblogs: Text Mining Study
JMIR Med Inform 2020;8(6):e17650
URL: http://medinform.jmir.org/2020/6/e17650/
doi:10.2196/17650
PMID:32574151

Original Paper

Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study

Weifeng Fu[1], PhD

[1] Liberal Arts College, Hunan Normal University, Changsha, China

Corresponding Author:
Weifeng Fu, PhD
Liberal Arts College
Hunan Normal University
36 Lushan Road
Changsha, 410081
China
Phone: 86 18973101748
Email: fwf1126@hunnu.edu.cn

Abstract

Background: Speech recognition is a technology that enables machines to understand human language.

Objective: In this study, speech recognition of isolated words from a small vocabulary was applied to the field of mental health counseling.

Methods: A software platform was used to establish a human-machine chat for psychological counselling. The software uses voice recognition technology to decode the user's voice information. The software system analyzes and processes the user's voice information according to many internal related databases, and then gives the user accurate feedback. For users who need psychological treatment, the system provides them with psychological education.

Results: The speech recognition system included features such as speech extraction, endpoint detection, feature value extraction, training data, and speech recognition.

Conclusions: The Hidden Markov Model was adopted, based on multithread programming under a VC2005 compilation environment, to realize the parallel operation of the algorithm and improve the efficiency of speech recognition. After the design was completed, simulation debugging was performed in the laboratory. The experimental results showed that the designed program met the basic requirements of a speech recognition system.

(JMIR Med Inform 2020;8(6):e18677) doi:10.2196/18677

KEYWORDS

speech recognition; isolated words; mental health; small vocabulary; HMM; hidden Markov model; programming

Introduction

Constraints on speech recognition such as small vocabularies, specific speakers, and isolated words need to be relaxed. At the same time, there are many new problems that must be solved. First, expanding the vocabulary makes it difficult to select and build templates. Second, in continuous speech, there is no obvious boundary between each phoneme, syllable, and word, and there is a phenomenon of coordinated pronunciation that is strongly influenced by the context of each pronunciation unit. Third, different people say the same words with different acoustic characteristics. Even when the same person speaks the same content multiple times, their physiological and psychological states may differ and cause notable differences in their speech. Fourth, there is often background noise or other interference accompanying speech. Therefore, the original template matching method is no longer applicable.

There have been further breakthroughs in using speech recognition technology for various applications for smartphones. This study focused on mental health issues and investigated the interaction between smartphone software and users' mental health based on speech recognition technology. This study involves basic application research on the use of intelligent software design and speech recognition technology in the context of mental health.

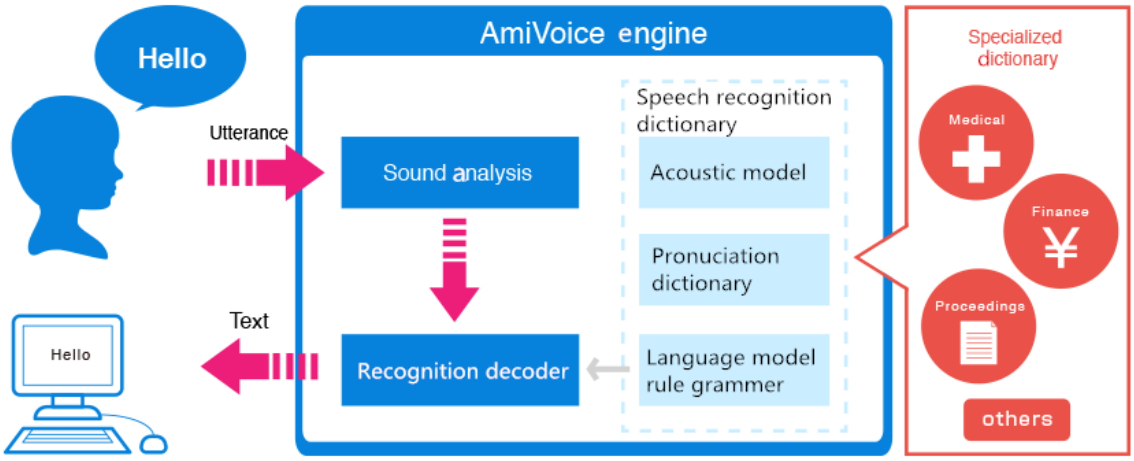
Methods

Programming of a Speech Recognition System Based on VC2005 Isolated Words

In this study, C language programming was used to implement data feature extraction based on the Markov model. It was then used to programmatically realize speech recognition for specific speech instances, as well as write speech recognition functions into functions that can be called by other modules. Additionally, it was used to implement a speech recognition system foundation, and to cultivate and improve the ability of the system to consult the literature and comprehensively use new knowledge [1].

Speech recognition is essentially a pattern recognition process, one by which an unknown speech pattern is compared with known reference patterns of speech, and the best-matched reference pattern is the recognition result. Figure 1 is a block diagram of an automatic speech recognition system based on the pattern matching principle [2].

Figure 1. Block diagram of speech recognition system.



Composition of an Isolated Word Speech Recognition System

The reference pattern is based on the template word unit shown in Figure 2. The main technical items of the isolated word speech recognition system are shown in Table 1.

Figure 2. References use templates as word units. HMM: Hidden Markov Model.

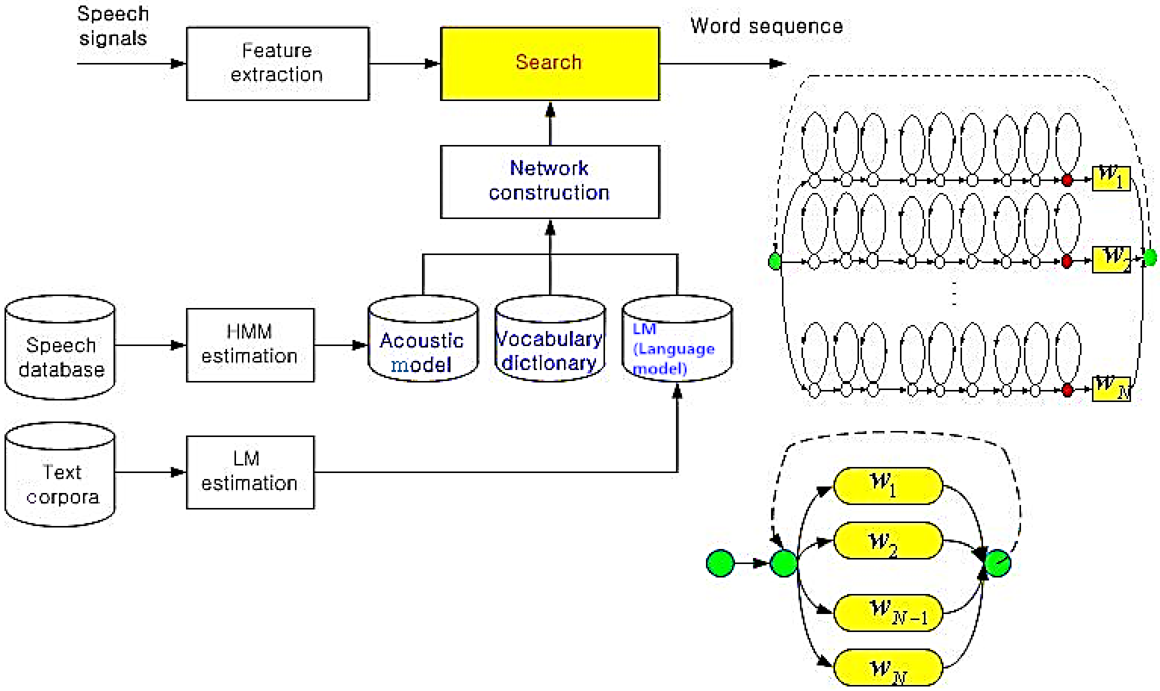


Table 1. Main technical items of the isolated word speech recognition system.

Technical items	Details
Vocabulary	Vocabulary fixed or variable, content (numbers, commands, place names, etc), acoustic similarity
Speaker	Specific speakers, nonspecific speakers
Generative method	Isolated vocalization, continuous vocalization
Analysis	Frequency domain analysis, cestrum domain analysis, linear prediction analysis
Mode change	Fixed or variable length, feature extraction, speech segmentation, factor recognition
Model approach	Multiple reference pattern matching method, statistical decision method, word formation recognition method
Standard mode	Standard template (multiple), word dictionary, probability distribution, generation rules
Input method	Phone-microphone (near microphone)
Vocal environment	Signal to noise ratio >30 decibels (dB)
Surroundings	Quiet office, spacious office, inside a moving car
Level	40-50 dB, 60-70 dB, 65-75 dB

Speech Recognition Design Process

Sample Voice Collection

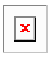
The standard Chinese numerals 0-9 were spoken and recorded indoors as a sample. The recording software used Microsoft Visual C++ Windows Media Player (Microsoft), with a sampling rate of 16 kHz and sampling bits of 16 bits. The voice data is stored in the .wav file format, and its audio format is Windows PCM (pulse-code modulation) [3].

Speech Signal Preprocessing

There were several elements involved in speech signal preprocessing. First, to digitize voice signals, data was extracted from the speech signal by sampling and quantizing. During data extraction, it is extremely important to master the storage form of the voice file, and to effectively extract and ascertain the meaning of each part of the data to improve the analysis of the data, and lay the groundwork for the next step.

Second, the high-frequency portion of the signal spectrum was enhanced and flattened, in order to facilitate channel parameter analysis or spectral analysis. Pre-emphasis of the speech signal is done by using the mean power spectrum and muzzle glottal excitation radiation effects; the high end at about 6 dB/octave is above 800 Hz, ie, 6 dB/octave (2 octaves) or 20 dB/decade (10 octaves). When seeking a voice signal spectrum, the higher the frequency, the smaller the corresponding component. For this reason, pre-emphasis is performed as part of preprocessing. The purpose of pre-emphasis is to flatten the signal spectrum, and hold the entire band from low to high frequency. The signal to noise ratio requirements can use the same spectrum or spectral analysis to analyze channel parameters. Pre-emphasis generally uses a first-order digital filter of the formula μ: $H(Z) = 1 - \mu z^{-1}$, where μ has a value close to 1, or formula $y(n) = x(n) - \alpha x(n-1)$, where $x(n)$ is the original signal sequence, $y(n)$ is the pre-emphasis sequence, and α is the pre-emphasis coefficient [4].

Third, preprocessing included endpoint detection and framed windowing. Breakpoint detection is mainly used to extract the effective part of the data. The threshold value is 0.3 (maximum value-minimum value). The speech signal is a typical nonstationary signal. In processing, a window function is generally used to intercept one segment for analysis. Part of the extracted signal is short-term stable. Another effect of windowing is to eliminate the Gibbs effect caused by truncation of infinite sequences. Common window functions [5] are as follows:



Both the Hamming window and the Haning window belong to the generalized raised cosine function. By analyzing their frequency response amplitude characteristics, it can be found that the rectangular window has good spectral smoothing performance, but the side lobe is too high, which may cause spectrum leakage and loss of high-frequency components. The Haning window decays too quickly and the low-pass characteristics are not smooth; the Hamming window is widely used because of its smooth low-pass characteristics and because it has the lowest side lobe height [6].

Mel Frequency Cepstral Coefficient Feature Representation

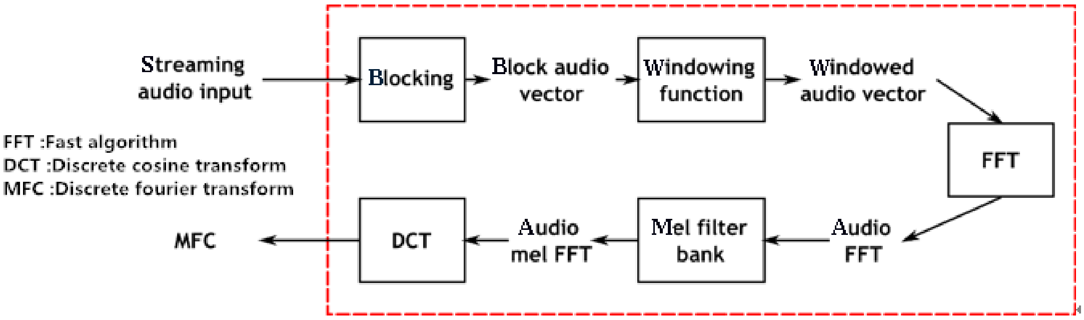
The training process of Mel Frequency Cepstral Coefficient (MFCC) parameters and Pearson Linear Correlation Coefficient (PLCC) parameters was extracted, that is, state transition matrix A, mixed Gaussian distribution weight matrix C, mean vector µ and covariance matrix U. A maximum likelihood estimation was performed.

MFCC Extraction

The human ear has different perception capabilities for speech at different frequencies; this is a nonlinear relationship. Combining the physiological structure of the human ear and using the logarithmic relationship to simulate the human ear's perception of speech at different frequencies, Davies and Merenstein proposed the concept of Mel frequency in 1980 [7]. The meaning of 1 Mel is 1/1000 of the tone perception degree of 1000 Hz. The conversion relationship between Hz frequency f_{Hz} and Mel frequency f_{Mel} is as follows:

The MFCC is proposed based on the above Mel frequency concept, and its computer flow is shown in Figure 3.

Figure 3. Mel Frequency Cepstral Coefficient (MFCC) calculation flow diagram. FFT: fast Fourier transform.



First, the original voice signal is pre-emphasized, and a frame of voice signal is obtained after frame-by-frame windowing. Second, the fast Fourier transform (FFT) is performed on a frame of speech signal to obtain the discrete power spectrum X (k) of the signal. Third, triangle filter center frequency f(m) and frequency response H (k) are calculated as follows:

In Equation 5, f_l and f_h are low-pass frequency filter bank coverage and high-pass frequency, respectively; F is the sampling frequency with the unit Hz; M is the number of filter bank filters; N represents the points that are FFT; B^{-1} is the inverse function of Equation 6.

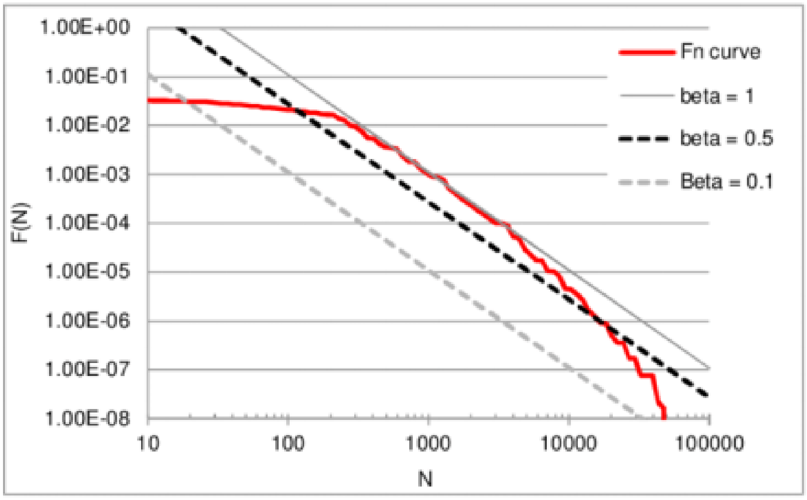
$B^{-1}(b) = 700(e^{b/127} – 1)$ **(6)**

Fourth, each filter produces an output spectral energy, after taking the number of coefficients so as to obtain the following set [8]:

A discrete cosine transform is used to convert S(m) to the time domain. The calculation process of the MFCC c(i) is as follows:

The curve and filter bank distribution corresponding to the MFCC's Hz-Mel scale are shown in Figure 4.

Figure 4. Mel Frequency Cepstral Coefficient (MFCC) scale corresponding curve.



HMM Pattern Matching

HMM pattern matching is a double random process evolved from Markov chains. An HMM with IV states is usually represented by λ = (A,B,π). The meaning of these parameters is explained as follows: N is the number of states of the model. An input observation sequence $O = o_1,o_2,\ldots o_T$ can only be in {S} at a certain moment, which is one of the N states of $\{S_1,S_2,\ldots,S_N\}$. A = $\{a_{ij}\}$ is the state transition probability matrix defined by the following equation: $a_{ij} = P(q_{t+1} = S_j \mid q_i = S_i)$, $1 \leq i, j \leq N$. It is an implicit Markov chain. The probability of each transition from state S_i to state S_j is only related to state S_i, and is not related to its previous state. The matrix elements must satisfy the following equation:

$\pi = \{\pi_1,\pi_2,\ldots,\pi_N\}$ is the initial probability distribution of each state, which represents the probability value that the observation sequence $O = o_1,o_2,\ldots o_T$ may be in each state of the model at ==========t = 1, that is, $p_i = P(q_1 = S_i)$, $i = 1,2,\ldots,N$, and it satisfies the following equation:

B is the output probability of any observation o_i in the input speech feature sequence $O = o_1, o_2, \ldots o_T$ in each state. It has two types: discrete and continuous. For the discrete HMM, B is a probability matrix $B = \{b_j(k)\}$, $j = 1,2,\ldots,M$; where $b_j(k) = P(o_k \mid q_t = S_j)$, and M is the total number of symbols in the coded symbol set and satisfies the following condition:

For the continuous HMM, $B = \{b_j(\mathrm{o})\}$, $1 \leq j \leq N$ and c_{ji}; among these, Jo is any feature vector K in the speech feature parameters, M is the number of Gaussian elements contained in each state, L is the weight of the jth state and the lth mixed Gaussian function, N is the normal Gaussian probability density function, m_{ji} represents the mean vector of the l mixed Gaussian element in the j state, and U_{ji} represents the covariance matrix of the l mixed Gaussian element in the j state, and it satisfies the following condition:

Results

Depending on different parameters of the HMM, it has different classification methods. One type of classification is to divide the HMM into two structures, ergodic and left to right, according to the transition probability matrix A = $\{a_{ij}\}$. The HMM experienced by each state is that any state in the model can reach all other states through a finite step; from left to right, the HMM increases with time, and the state serial number is nondecreasing. This model is divided into spanning and no spanning. The HMMs of various states are mostly used for speaker recognition, language recognition, etc. The content of speech has a strong correlation with timing. This timing can be expressed by the state relationship, so speech recognition must use the left to right HMM structure. This study is based on isolated word speech recognition, and it is not allowed to skip a certain part of the middle of a speech fragment, so the HMM structure of left to right without crossing must be used. Its state transition probability matrix A = $\{a_{ij}\}$ must satisfy $a_{ij} = 0$, $j \neq i$ and $j \neq i + 1$ [9].

Another classification method is to divide HMMs into continuous, discrete, and semicontinuous based on different output probabilities B. The output probability B of each state of the discrete HMM is a discrete probability matrix, and the vector of the feature parameter of the speech signal must be vector quantized before use. The output probability B of the continuous HMM is a continuous output probability density function. It has three forms: single, mixed, and differentiated Gaussian probability density function. The semicontinuous HMM is a method that combines discrete HMM and continuous HMM. This paper uses a continuous HMM.

The following problems are to be solved by the isolated word speech recognition system based on the HMM: First, how to determine an optimal state transition sequence $q = (q_1, q_2, \ldots, q_T)$, and calculate the output probability P(O|λ) of the observation sequence $O = o_1, o_2, \ldots o_T$ to the HMM, and judge the recognition result of the voice command based on this probability. Second, how to adjust the parameters that λ = (A,B,π) to maximize the output probability P(O|λ). This is a problem of parameter training of the HMM. In the process of solving the above two problems, the output probability needs to be calculated, which is another key problem that needs to be solved by this algorithm [10].

Discussion

Small Vocabulary Speech Recognition System Applied in the Field of Mental Health

Speech Recognition System and Acquisition Method

For different speech types that need to be recognized [10], the system collected data in different ways. For mobile phone software, the intelligent degree of speech recognition is completely dependent on the preset scheme. The same speaker's speech may get completely different results due to different collection methods preset by the recognition system. Therefore, for users with special voice types, the mobile phone software adopts multiple (1-3) collection methods to reduce errors.

Speech Processing System

The speech processing system mainly analyzes and processes speech to achieve the purposes of transmission, automatic recognition, and machine understanding. The analysis and processing are implemented based on the filtering, sampling, and Fourier transform algorithms; the mobile phone software runs the experimental results. The speech processing system also processes voice signals such as echo, user's voice disturbance, and voice noise to manage some typical voice transmission problems.

Establishment of Related Databases

A psychological database that contains psychological cases and current user psychological data was established. It establishes a relationship between all data in the database and uses the data dictionary to expand the function of the table to make the database design simpler. The database also needs to regularly update relevant information to better enable the software platform to provide users with mental health information. The user steps are as follows: (1) After opening the mobile phone software, the system prompts the user to fill in relevant information such as gender and age (personal information). (2) The voice chat system will conduct a human-machine voice chat, with humorous and interesting content occasionally mixed with some questions. (3) After the chat is over, the user is notified that there is a waiting time. The software system analyzes the voice chat data and further analyzes the experimental results. (4) The user is notified of the analysis result, and the software performs the first operation on the user if they have identified psychological problems. (5) The software then establishes a specific personal psychological treatment plan for users with mental disorders.

Application of Analysis Software

The experimental data showed that our mobile phone mental health software meets the requirements for accuracy, practicability, and simplicity. The software was able to realize specific operations on related data by programming, to obtain the most reliable parameters and achieve an accurate probability of the user's voice information, thereby inferring any psychological changes. The program was able to make a scientific, professional, and safe analysis of users' mental health with different personality characteristics. Using this software is convenient for users.

Conclusion

In response to the special requirements of speech recognition, the design of this software system is based on digital signal processing and uses a fast Fourier transform. Overall, the design requirements were met. However, due to time and knowledge limitations, there are still existing problems with the design, such as the incomplete treatment of environmental noise effects. There is room for improvement in this software system. This article introduces this research and factual issues such as the application of mobile phone mental health software. The software platform is quantified and modularized using user needs. It analyzes and processes specific experimental data, emphasizing that mental health software in a mobile phone is convenient.

Acknowledgments

The author would like to thank Dr Yingmei Xu of the faculty from Psychological Counseling and Treatment Center of Peking University Hospital for her help.

Conflicts of Interest

None declared.

References

1. Boussaid L, Hassine M. Arabic isolated word recognition system using hybrid feature extraction techniques and neural network. Int J Speech Technol 2017 Nov 23;21(1):29-37. [doi: 10.1007/s10772-017-9480-7]
2. Bennettlevy J, Martin E, Bridgman H, Carey T, Isaacs AN, Little F. Mental health academics in rural and remote Australia. Rural and remote health 2016;16(3):3793-3793.
3. Wu S. A Traffic Motion Object Extraction Algorithm. Int J Bifurcation Chaos 2016 Jan 14;25(14):1540039. [doi: 10.1142/s0218127415400398]
4. Fujii Y, Fujii K, Yoon J, Sugahara H, Kitano N, Okura T. The Effects Of Low-intensity Exercise On Depressive Symptoms In Socially-isolated Older Adults. Medicine & Science in Sports & Exercise 2016;48:1052. [doi: 10.1249/01.mss.0000488166.06405.c1]
5. Elovainio M, Hakulinen C, Pulkki-Råback L, Virtanen M, Josefsson K, Jokela M, et al. Contribution of risk factors to excess mortality in isolated and lonely individuals: an analysis of data from the UK Biobank cohort study. The Lancet Public Health 2017 Jun;2(6):e260-e266. [doi: 10.1016/s2468-2667(17)30075-0]
6. 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]
7. Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions On Acoustics, Speech and Signal Processing 1980:357-366.
8. Banbury A, Nancarrow S, Dart J, Gray L, Parkinson L. Telehealth Interventions Delivering Home-based Support Group Videoconferencing: Systematic Review. J Med Internet Res 2018 Feb 02;20(2):e25. [doi: 10.2196/jmir.8090]
9. R.P K. Cognitive deterioration following intracranial haemorrhage: Predominantly dependent on cognitive health before the event. Nederlands tijdschrift voor geneeskunde 2016;160(21):9697-9697.
10. 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]

Abbreviations

dB: decibel
FFT: fast Fourier transform
HMM: Hidden Markov Model
MFCC: Mel Frequency Cepstral Coefficient
PLCC: Pearson Linear Correlation Coefficient

Edited by K Kalemaki; submitted 12.03.20; peer-reviewed by J Firas, X Liu, Y Cao; comments to author 19.03.20; revised version received 21.03.20; accepted 21.03.20; published 03.06.20.

Please cite as:
Fu W
Application of an Isolated Word Speech Recognition System in the Field of Mental Health Consultation: Development and Usability Study
JMIR Med Inform 2020;8(6):e18677
URL: https://medinform.jmir.org/2020/6/e18677
doi:10.2196/18677
PMID:32384054

Original Paper

Medical Emergency Resource Allocation Model in Large-Scale Emergencies Based on Artificial Intelligence: Algorithm Development

Lin Du[1], MSc

[1] School of Information Science and Engineering, Qilu Normal University, Jinan, China

Corresponding Author:
Lin Du, MSc
School of Information Science and Engineering
Qilu Normal University
No 33, Shanshi East Road
Jinan
China
Phone: 86 13793161610
Email: dul1028@163.com

Abstract

Background: Before major emergencies occur, the government needs to prepare various emergency supplies in advance. To do this, it should consider the coordinated storage of different types of materials while ensuring that emergency materials are not missed or superfluous.

Objective: This paper aims to improve the dispatch and transportation efficiency of emergency materials under a model in which the government makes full use of Internet of Things technology and artificial intelligence technology.

Methods: The paper established a model for emergency material preparation and dispatch based on queueing theory and further established a workflow system for emergency material preparation, dispatch, and transportation based on a Petri net, resulting in a highly efficient emergency material preparation and dispatch simulation system framework.

Results: A decision support platform was designed to integrate all the algorithms and principles proposed.

Conclusions: The resulting framework can effectively coordinate the workflow of emergency material preparation and dispatch, helping to shorten the total time of emergency material preparation, dispatch, and transportation.

(JMIR Med Inform 2020;8(6):e19202) doi:10.2196/19202

KEYWORDS

medical emergency; resource allocation model; distribution model; large-scale emergencies; artificial intelligence

Introduction

After an emergency, enough emergency supplies should be sent to the disaster area as soon as possible to minimize the casualties. These emergency supplies include medical equipment, medicine, food, drinking water, and protection from the cold. An important factor that affects the length of time needed to deliver emergency supplies is the amount of advance preparation. If most of the emergency supplies are already prepared in peacetime, they can be directly transported to the disaster area when the emergency arrives, saving time in accumulating supplies. The government's work in the preparation of emergency supplies generally follows a particular process. First, the government predicts possible future emergency needs based on relevant information such as the type, time, location, intensity, and probability of future emergency events, which is provided by relevant scientific research institutions. Then, it obtains funds for emergency materials through various methods. Next, it contacts relevant companies and purchases emergency materials. Finally, it stores the emergency materials that have been purchased. The government's daily preparation of emergency supplies is not without cost, and it should consider the two goals of ensuring safety and reducing expenditure.

Therefore, the first purpose of this paper is to solve the problem of the type and quantity of emergency materials prepared in advance. The second purpose of this paper is to develop an internet-based decision support platform that can perform that function. Whether the government is judging the types and quantities of emergency materials demanded or researching emergency material scheduling strategies, it must rely on experts in related fields, including geological experts, astronomical experts, and other experts related to the nature of the emergency. It also includes medical experts and emergency management experts who need to be involved regardless of the type of emergency. Gathering these experts and government decision makers in the same place in a short time is basically impossible, and there is no need to do so to carry out decision-making work. The use of the internet can solve this problem by achieving a virtual assembly of experts and government decision makers.

Methods

Preparation, Dispatch, and Transportation of Emergency Materials for Major Emergencies

Coordinated Operation of Emergency Material Preparation and Transportation

After determining the need and demand for the type of emergency supplies, the shipping process begins. In major emergencies, the emergency supply–shipping process consists of 3 subprocesses: preparation of emergency supplies, scheduling of emergency supplies, and transportation of emergency supplies. First, when the government receives a new demand for emergency supplies, the government emergency response decision-making group needs to prepare these emergency supplies through various channels. Then, it convenes an expert group to analyze the current traffic information, material transport routes, and expected transit time, and it uses computers to calculate optimal vehicle transportation routes. Finally, vehicles carry emergency supplies to the affected area and in accordance with a predetermined route.

The transportation of emergency supplies also has 2 subprocesses: the first is the preshipment of emergency supplies, and the second is the shipment of emergency supplies once the schedule has been finalized. After the government demands preparation of emergency supplies, the staff responsible immediately begin to prepare the supplies without having to consider the progress of the other subprocesses. As a result, the preparation subprocess in not the entire critical path.

In the process of preparing, scheduling, and transporting medical supplies, the subprocesses of preparing emergency supplies, scheduling their delivery, and transporting them are carried out simultaneously. Because both the distance between the hub and each disaster and the number of vehicles involved in the transport are already determined, the optimal transit times of emergency supplies cannot be changed [1].

Subprocess of Emergency Material Preparation for Major Emergencies

According to the previous analysis, in major emergencies there are 3 main channels for the government to prepare emergency materials. First, the government directly dispatches the prepared emergency materials in stock. Second, the government supplements and purchases emergency materials from suppliers. Third, the government asks the media to request that the public donate emergency supplies.

In the first channel, the cost of accessing the stocks of emergency materials is much less than the cost of urgently sourcing the same emergency materials from suppliers. In addition, after receiving the transfer notification regarding the stock, the government can immediately ship the emergency materials without spending time in other areas. Finally, the place of storage for the emergency supply stocks is either the distribution center or very close to the distribution center, so the transportation time is also much shorter than that of the other material preparation channels. However, if the emergency supplies in stock cannot meet the rescue needs, the government can use the 2 other channels to prepare emergency supplies. The most widely used channel is ordering emergency materials from suppliers. Requesting that the public donate emergency supplies is relatively rare. Because the time of arrival, quantity, and quality of emergency materials obtained in this way cannot be controlled by the government, they cannot be included in the unified planning.

To sum up, these 3 channels together comprise the subprocess of emergency material preparation for major emergencies.

Subprocess of Emergency Material Dispatch for Major Emergencies

There are 2 key steps that need to be completed in order to dispatch emergency supplies. The first step is expert evaluation, and the second step is computer calculation [2].

First, the experts comprehensively evaluate real-time road conditions based on aerial photographs, information provided by local governments, and feedback information from drivers of transportation vehicles, thereby determining the best route between disaster sites, as well as the required transportation time. The expert's evaluation is very important because the result of the evaluation determines the optimal transportation strategy. However, in order to ensure the continuity of the entire process, it is necessary to strictly limit the time used by expert evaluation. For a cluster decision, which is made by multiple experts independently, experts generally make one or more rounds of judgment before making a final decision. Under this scenario, it obviously saves more time if experts make only 1 round of judgment. Since each expert judges the existing information differently, it is possible for them to evaluate each possibility of road conditions. In order to make the conclusion drawn by the experts conform to a unified standard, the increase in the expected transportation time between 2 affected areas compared with original transportation time between them can be set to m possible minutes. Then experts can evaluate the probability of these transportation times. If there are n experts participating in the evaluation, the probability of the i expert evaluating the m cases is $P_i(L_1)$,$P_i(L_2)$,…$P_i(L_m)$. According to the evidence synthesis algorithm, the expert's evaluation of the probability of the j case is:

See Equation 2:

As a result, the latest expected increase in time for the vehicle to travel a certain route is the situation with the highest evaluation probability. Immediately afterwards, the latest traffic information is substituted into the material transportation scheduling model proposed above and the optimal transportation strategy can be calculated. This calculation process is completed by computer [3].

Models for Emergency Material Preparation, Dispatch, and Transportation

Workflow System for Emergency Material Preparation, Dispatch, and Transportation

If the time for vehicle k + 1 and vehicle k to return to the distribution center is $t_{(k+1)}$ and t_k, respectively, and vehicle k + 1 returns after vehicle k, and if there is no other vehicle between the 2 vehicles, then the time interval is $t_{(k+1)} - t_k$. t_p refers to the transportation of materials. If the total number of vehicles participating in transportation is m, the minimum time interval for all returning vehicles is:

$$t_p = \min\,[t_2 - t_1,\ldots t_{(k+1)} - t_k,\ldots t_m - t_{(m-1)}]\ \textbf{(3)}$$

The subprocesses for emergency supply preparation and emergency supply scheduling are together called the material preparation and scheduling process. Each batch of materials must go through this material preparation and scheduling process in order to be transported to the disaster area. If there is a failure, the government must look for other suppliers to enter the next round of negotiations. This process refers to the time period required for this meeting [4].

Here, t_{TN} refers to the required time, and E refers to the set of optimal solutions. This formula is to calculate the optimal solution in the shortest time:

The emergency materials prestored in the government warehouse can be shipped immediately after getting the transportation instructions. The preparation time is 0. The time-consuming part of the material preparation subprocess is the following: the government directly dispatches the stock emergency materials prepared in advance, the government supplements and purchases emergency materials from suppliers (if necessary), and the government requests the society to donate emergency materials through the media (if necessary). The government waits for the maximum value of the three processes. The total running time of this material preparation and scheduling process is shown below:

In this equation, t_{TN} is the time used by the government to negotiate with the supplier, t_{TP} is the time spent by the supplier to prepare materials after the negotiation is successful, t_{TM} is the time required for the media to publish a report, and t_{TD} is the time between the public seeing the media report and preparing to donate. Regarding the time to receive materials, t_{TL} is the average time for emergency materials to be transported from the supplier to the distribution center, t_{TE} is the expert's estimated transit time of each transportation route, and t_{TA} is the computer calculation using the emergency material transportation scheduling model.

According to the previous analysis, this material preparation and dispatch process should be completed within the minimum time interval between the 2 transport vehicles arriving at the material distribution center:

$$t_p = t'_p\ \textbf{(6)}$$

Set in the period of [0,T_1], where T_1 equals the first time, the time of occurrence of each demand for emergency supplies is independent and follows a negative exponential distribution. During this period, when urgent materials are needed, we can know the average arrival rate of demands for emergency supplies:

The average cost of demands for emergency supplies is

According to the principle of queueing theory, the length of stay follows a negative exponential distribution with parameter $\mu-\lambda$. Then, the probability density of the dwell time w is:

The average length of stay can be determined as follows, where w is the final result:

The average waiting time can be found below:

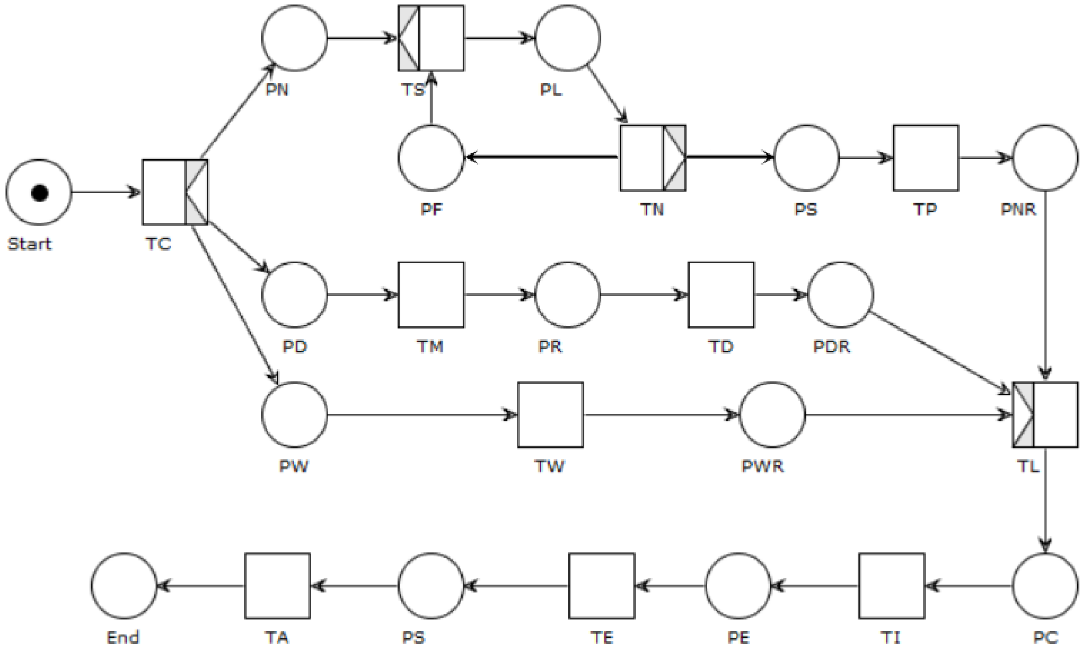
In addition, emergency materials, such as medical supplies and living necessities, must be delivered within a certain period of time. Otherwise, it will cause greater casualties. To prevent this, the formula can be set so that these emergency materials must be delivered within the [0,T_1] period. Among the formulas, suppose that a certain link i in the material preparation and scheduling process needs n staff to participate together, and the amount of time of this link is t_{pi}. The total effective working time of the n staff members is nt_{pi}. c_{pi} refers to a single specific index and R_p refers to the composite index. If the time utilization rate of these staff is set to $r_p(0<r_p<1)$, then the number of staff required is

Using the Whooped (Microsoft Corp) software to conduct a static test on the system, the system has passed the coverage, boundedness, and activity tests, so the design of this system is reasonable [5].

Simulation System Framework for Emergency Material Preparation and Dispatch

Based on the above ideas, this paper establishes a simulation system framework for emergency material preparation and dispatch, as shown in Figure 1. The simulation system framework presents all links in the material preparation and material transportation subprocesses in the first-level system, ignoring the difference in the sourcing and preparation time of each batch of emergency materials and removing all emergency materials in the government emergency process [6].

Figure 1. Framework of emergency material preparation and dispatch simulation system. PC: PC-BSD installer; PD: product department; PDR: project definition report find suppliers; PE: portable executable; PF: planned finish date; PL: pay local; PN: part number; PNR: prior notice required; PR: pattern recognition; PS: performance standard; PW: present worth; PWR: process work request; TA: computer computes optimal transportation routes; TC: trace cache; TD: corporate donations of emergency supplies; TE: experts assess the trip transport routes by the expected time; TI: the government gives the current traffic information; TL: set off emergency supplies destined for distribution center; TM: media reports; TN: government negotiations with suppliers; TP: suppliers prepare emergency supplies; TS: find suppliers; TW: government warehouse to prepare emergency supplies.



Results

Radio frequency identification technology can be mainly applied to the storage of emergency supplies. It is connected to a database, and its functions include the storage and inventory management of emergency materials, as well as the management of emergency materials after major emergencies. In daily situations, the government puts the emergency materials into storage after purchasing them. During this time, the radio frequency identification technology can automatically identify and store the materials according to their category and storage environment requirements. Additionally, the long-term storage of materials is generally limited by shelf life. This technology can automatically pick out materials that have passed their shelf life and prompt management personnel to replace them with new materials. Moreover, as the amount of emergency materials stored changes, the material storage warehouse may also need to adjust its structure or expand its scale. At this time, the technology can help managers find the materials that need to be transferred and ensure the orderly progress of the entire transfer [7].

After major emergencies occur, emergency supplies need to be transported out of the warehouse to the affected area. Radio frequency identification technology can enable managers to track the status of each emergency item at any time and help them formulate the best transportation plan. More importantly, the use of this technology allows managers to more quickly calculate the difference between the inventory and the demand for various emergency supplies.

The application of the Internet of Things technology in the process of emergency material preparation and transportation in major emergencies is mainly achieved through multiple terminal devices of different types, such as active and passive microwave transmissions, reception devices involved in radio frequency identification technology, GPS terminal signal receivers, ground control instruments, and satellites. The most important role played by these terminal devices is to facilitate the collection of multiple types of decision data. However, in the government emergency process developed in this paper, a large amount of data entry is required. Therefore, the combination of the decision support platform with the Internet of Things technology is feasible and even efficient.

Based on this idea, this paper proposes a framework for the organic integration of the related technologies of the Internet of Things and the government emergency process [8]. After the integration, the new decision support platform will effectively integrate emergency supplies, rescuers, decision makers, and computer artificial intelligence into a large system, which will significantly increase the speed of information transmission and processing, thereby greatly improving the operational efficiency of preparation, scheduling, and transportation of emergency supplies [9].

Discussion

This paper establishes a model for emergency material preparation and dispatch based on queueing theory; further establishes a workflow system for emergency material preparation, dispatch, and transportation based on a Petri net; and provides a simplified simulation system framework for emergency material preparation and dispatch with high operating efficiency. It can effectively coordinate the workflow of emergency material preparation and dispatch, shortening the total time needed by these processes. Based on the Internet of Things technology that can be used in the process of emergency material transportation, an integrated framework for emergency material financing is constructed. This paper also provides an interface to the model solution software on the decision support platform so that this decision support platform integrates all the principles and algorithms of emergency material financing and transportation proposed in this paper.

Acknowledgments

The author acknowledges the Key Research and Development Program of Shandong Province (Grant No. 2019GSF108010).

Conflicts of Interest

None declared.

References

1. Tucci VT, Moukaddam N, Alam A, Rachal J. Emergency Department Medical Clearance of Patients with Psychiatric or Behavioral Emergencies, Part 1. Psychiatr Clin North Am 2017 Sep;40(3):411-423. [doi: 10.1016/j.psc.2017.04.001] [Medline: 28800798]
2. Eaton-Evans T. Managing medical emergencies at endurance rides. In Practice 2019 Jul 11;41(6):270-274. [doi: 10.1136/inp.l4108]
3. Zhou QS, Olsen TL. Inventory rotation of medical supplies for emergency response. European Journal of Operational Research 2017 Mar;257(3):810-821. [doi: 10.1016/j.ejor.2016.08.010]
4. Ki M. Surveillance and epidemiologic investigation in public health emergencies caused by infectious diseases. J Korean Med Assoc 2017;60(4):292. [doi: 10.5124/jkma.2017.60.4.292]
5. Schönfeldt-Lecuona C, Gahr M, Schütz S, Lang D, Pajonk FGB, Connemann BJ, et al. Psychiatric Emergencies in the Preclinical Emergency Medicine Service in Ulm, Germany in 2000 and 2010, and Practical Consequences. Fortschr Neurol Psychiatr 2017 Jul;85(7):400-409. [doi: 10.1055/s-0042-122709] [Medline: 28768348]
6. Nable JV, Tupe CL, Gehle BD, Brady WJ. Is there a doctor on board? In-flight medical emergencies. Cleve Clin J Med 2017 Jun;84(6):457-462 [FREE Full text] [doi: 10.3949/ccjm.84a.16072] [Medline: 28628427]
7. Albelaihi H, Alweneen A, Ettish A, Alshahrani F. Knowledge, Attitude, and Perceived Confidence in the Management of Medical Emergencies in the Dental Office: A Survey among the Dental Students and Interns. J Int Soc Prev Community Dent 2017;7(6):364-369 [FREE Full text] [doi: 10.4103/jispcd.JISPCD_414_17] [Medline: 29387622]
8. Aftergood DE. In-Flight Medical Emergencies. N Engl J Med 2016 Jan 21;374(3):292. [doi: 10.1056/NEJMc1512716] [Medline: 26789896]
9. Goertzel B. Artificial general intelligence. Lecture Notes in Computer Science 2017;56(2):32-39. [doi: 10.1007/978-3-540-68677-4]

Edited by H Li, Z Du, K Kalemaki; submitted 07.04.20; peer-reviewed by Z Liu, Y Liu, M Zhang; comments to author 15.04.20; revised version received 22.04.20; accepted 23.04.20; published 25.06.20.

Please cite as:
Du L
Medical Emergency Resource Allocation Model in Large-Scale Emergencies Based on Artificial Intelligence: Algorithm Development
JMIR Med Inform 2020;8(6):e19202
URL: http://medinform.jmir.org/2020/6/e19202/
doi:10.2196/19202
PMID:32584262

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org