JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 10 (2022), Issue 4 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Corrigenda and Addendas

Review

The Use of Artificial Intelligence–Based Conversational Agents (Chatbots) for Weight Loss: Scoping Review and Practical Recommendations

Han Shi Jocelyn Chew[1], BSN, PhD

Alice Lee Centre for Nursing Studies, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Corresponding Author:
Han Shi Jocelyn Chew, BSN, PhD
Alice Lee Centre for Nursing Studies
Yong Loo Lin School of Medicine
National University of Singapore
Level 3, Clinical Research Centre, Block MD11
10 Medical Drive
Singapore, 117597
Singapore
Phone: 65 65168687
Email: jocelyn.chew.hs@nus.edu.sg

Abstract

Background: Overweight and obesity have now reached a state of a pandemic despite the clinical and commercial programs available. Artificial intelligence (AI) chatbots have a strong potential in optimizing such programs for weight loss.

Objective: This study aimed to review AI chatbot use cases for weight loss and to identify the essential components for prolonging user engagement.

Methods: A scoping review was conducted using the 5-stage framework by Arksey and O'Malley. Articles were searched across nine electronic databases (ACM Digital Library, CINAHL, Cochrane Central, Embase, IEEE Xplore, PsycINFO, PubMed, Scopus, and Web of Science) until July 9, 2021. Gray literature, reference lists, and Google Scholar were also searched.

Results: A total of 23 studies with 2231 participants were included and evaluated in this review. Most studies (8/23, 35%) focused on using AI chatbots to promote both a healthy diet and exercise, 13% (3/23) of the studies used AI chatbots solely for lifestyle data collection and obesity risk assessment whereas only 4% (1/23) of the studies focused on promoting a combination of a healthy diet, exercise, and stress management. In total, 48% (11/23) of the studies used only text-based AI chatbots, 52% (12/23) operationalized AI chatbots through smartphones, and 39% (9/23) integrated data collected through fitness wearables or Internet of Things appliances. The core functions of AI chatbots were to provide personalized recommendations (20/23, 87%), motivational messages (18/23, 78%), gamification (6/23, 26%), and emotional support (6/23, 26%). Study participants who experienced speech- and augmented reality–based chatbot interactions in addition to text-based chatbot interactions reported higher user engagement because of the convenience of hands-free interactions. Enabling conversations through multiple platforms (eg, SMS text messaging, Slack, Telegram, Signal, WhatsApp, or Facebook Messenger) and devices (eg, laptops, Google Home, and Amazon Alexa) was reported to increase user engagement. The human semblance of chatbots through verbal and nonverbal cues improved user engagement through interactivity and empathy. Other techniques used in text-based chatbots included personally and culturally appropriate colloquial tones and content; emojis that emulate human emotional expressions; positively framed words; citations of credible information sources; personification; validation; and the provision of real-time, fast, and reliable recommendations. Prevailing issues included privacy; accountability; user burden; and interoperability with other databases, third-party applications, social media platforms, devices, and appliances.

Conclusions: AI chatbots should be designed to be human-like, personalized, contextualized, immersive, and enjoyable to enhance user experience, engagement, behavior change, and weight loss. These require the integration of health metrics (eg, based on self-reports and wearable trackers), personality and preferences (eg, based on goal achievements), circumstantial behaviors (eg, trigger-based overconsumption), and emotional states (eg, chatbot conversations and wearable stress detectors) to deliver personalized and effective recommendations for weight loss.

(JMIR Med Inform 2022;10(4):e32578) doi:10.2196/32578

KEYWORDS

chatbot; conversational agent; artificial intelligence; weight loss; obesity; overweight; natural language processing; sentiment analysis; machine learning; behavior change; mobile phone

Introduction

Background

The global prevalence of obesity has risen dramatically over the past 50 years and has now reached a state of a pandemic [1]. It was estimated that approximately 39% of the global adult population and more than 18% of the younger population were overweight in 2016 [2]. This creates a pressing public health concern because overweight and obesity increase one's risk of disabilities, morbidities, and mortality from cardiometabolic diseases (eg, coronary artery disease and diabetes mellitus) [3], musculoskeletal disorders [4], cancers [5], and communicable diseases [6]. Having a high BMI have also been associated with a 32% increase in the likelihood of developing depression than having a normal weight, lowering one's quality of life [7,8]. Although the prevalence of overweight and obesity is higher in adults, a meta-analysis reported that children and adolescents with obesity had a 5 times higher risk of transitioning to adulthood with obesity [9]. This highlights the importance of targeting both the adult and younger population in global weight management efforts.

Besides the minority cases where overweight and obesity are caused by pharmacological, metabolic, or genetic etiologies, people enrolled in weight loss programs are often prescribe diet (that reduces calorie intake) and exercise (that increases calorie expenditure) plans that create a state of prolonged calorie deficit. However, a major challenge of such interventions is the lack of adherence to restrictive lifestyle plans, often due to a lack of motivation and self-control (ie, cognitive inhibition: ability to control impulses) [10]. To overcome such challenges, health coaching has been shown to enhance the initiation and sustainability of weight loss efforts through nutrition and exercise education, goal setting, periodic progress monitoring, and positive encouragement [11]. However, such programs are labor intensive and resource inefficient [11,12]. Brief counseling techniques such as motivational interviewing have also been shown to improve one's lifestyle behaviors but multiple empirical studies and systematic reviews have reported no significant superiority in interventional effectiveness when compared with other active comparators such as health coaching [13,14]. The findings were regardless of age and the mode of delivery [13-16], suggesting that current interventions are effective but impeded by their resource intensiveness (eg, time, manpower, and infrastructure) for coach training, program implementation, coordination, maintenance, and sustenance.

Recent technological advancements have enabled the use of computerized chatbots, also known as conversational agents (CAs), to mimic the role of human health coaches. Although terms such as chatbots, conversational artificial intelligence (AI), intelligence chatbots, and CAs are often used interchangeably, chatbots can be distinguished as those with and without AI [17]. In this paper, chatbots refer to computer software that is capable of having a conversation with someone and AI refers to the machinery mimicry of human intelligence to perform human tasks such as decision-making and problem solving, largely using machine learning [18]. Traditional rule-based chatbots without AI are only capable of identifying a limited number of client intents based on utterance interpretation of specific keywords [17]. This limits the degree of human conversation mimicry and hence the number of meaningful conversational turns to establish a motivational human-chatbot rapport. In contrast, AI chatbots are capable of machine learning to understand human intents and sentiments, thereby conversing with human-like demeanors to enhance human-chatbot interactions. This requires the use of natural language processing (NLP) for use cases such as natural language inference, sentiment analysis, and questioning and answering. In recent years, NLP has advanced from using traditional recurrent neural network models that analyze short texts for tasks such as summarization, translation, and abstraction to pretrained transformer models that analyze long texts as a whole to perform higher-level tasks of understanding and contextualization. Recent transformer models include Bidirectional Encoder Representations from Transformers by Google [19], Generative Pretrained Transformer (GPT-2 [20] and GPT-3 [21]) by Open AI, XLNet [22], and Turing Natural Language Generation by Microsoft [23]. The use of such technology in chatbots is more intuitive and able to express human emotions or cognitive responses such as empathy to enhance social presence, human-machine trust, emotional bond, user acceptability, and engagement [24]. A popular NLP platform used to develop and deploy such chatbots is Dialogflow, a user-friendly Google cloud-based platform capable of deploying text- and speech-based chatbots on various smartphone apps, websites, and Internet of Things (IoT) devices and appliances.

The use of AI in weight loss has been widely studied for its ability to efficiently and intuitively track diet, exercise, and energy balance. However, less is known about its ability to provide effective recommendations and behavioral nudges to enhance weight loss success [18]. Chatbots possess great potential as a communication vector for behavioral nudges through a sustained period of health coaching, thereby supplementing the role of a human health care professional in monitoring and counseling for weight loss. In addition, chatbots can provide 24/7 real-time monitoring, on-demand counseling, and personalized recommendation services conveniently through one's preferred device and social communication platform (eg, WhatsApp, Telegram, and Facebook Messenger). Such functions have been shown to increase usability, user acceptability, engagement, and potential weight loss success because of their convenience and instantaneousness [25]. However, this is contingent upon the ability to forge a human-like rapport with users, which is one of the largest challenges in chatbot development. Moreover, little is known about the chatbots that

have been developed to address health issues that require multiple long-term behavior changes such as for overweight and obesity [26].

Objectives

This study aims to provide an overview of the potential use of AI chatbots for weight loss in people with overweight and obesity, and identify the essential components to prolong user engagement in AI chatbot–delivered weight loss programs. The term chatbot will hitherto refer to AI chatbots unless otherwise stated.

Methods

This scoping review was performed according to the 5-stage framework by Arksey and O'Malley [27] and reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Multimedia Appendix 1) [28].

Stage 1: Identifying the Research Question

The research question for this study was developed based on the population, intervention, comparison, and outcomes framework, *What is known about the potential use of AI chatbots for weight loss in people with overweight and obesity and how can we prolong user engagement in AI chatbot–delivered weight loss programs?*

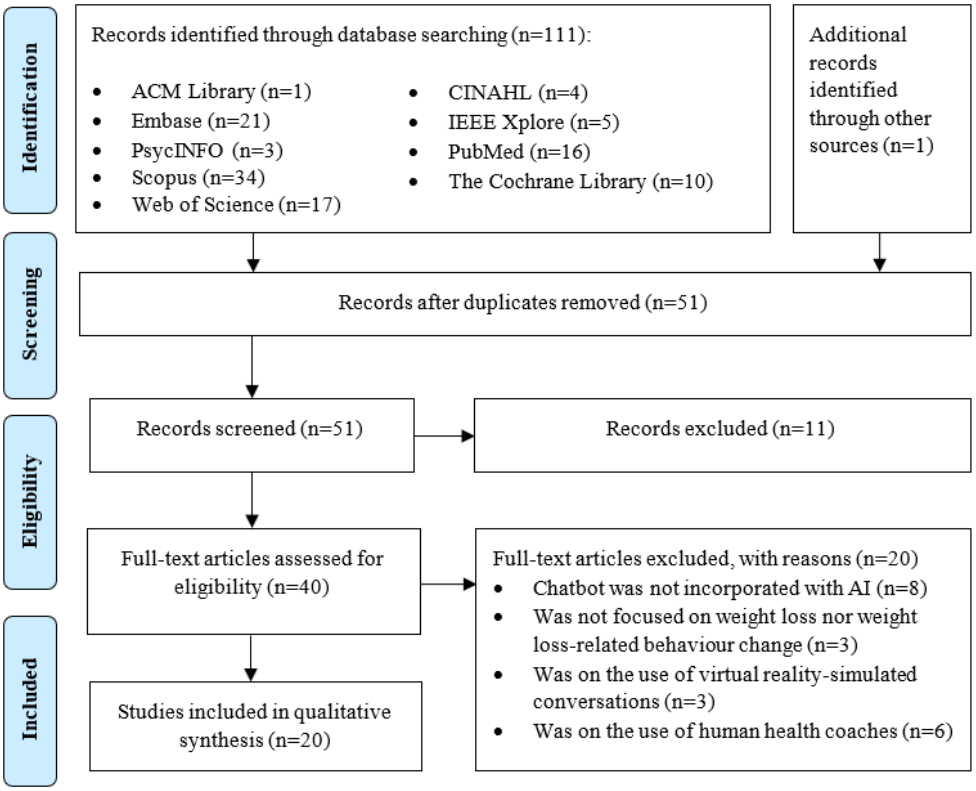
Stage 2: Identifying Relevant Studies

The Cochrane Database of Systematic Reviews and PROSPERO databases were first searched to confirm that there was no previous systematic review on this topic. All studies published until July 9, 2021, were searched across nine databases: ACM Digital Library, CINAHL, Cochrane Central, Embase, IEEE Xplore, PsycINFO, PubMed, Scopus, and Web of Science. Keywords were permuted by iterative searching of PubMed and Medical Subject Headings terms using initial terms such as *chatbot* and *obesity*. The final search terms used were *overweight*, *obes**, *chatbot**, *conversational agent**, *virtual coach**, *artificial intelligence*, *machine learning*, and *health coach**. The search strings connected using the Boolean operators are detailed in Multimedia Appendix 2. To ensure a comprehensive and extensive search on this topic, gray databases such as arXIV, Mednar, ProQuest Dissertation and Theses Global, and Science.gov were searched. Additional articles were also hand-searched from the reference lists of the included studies and the first 10 pages of Google Scholar.

Stage 3: Study Selection

The eligibility criteria for article inclusion were decided post hoc after an iterative screening of the resultant titles and abstracts and deeper familiarity with the topic. Articles that focused on the use of AI-based chatbots for weight loss were included. Given the lack of studies that focused on the use of AI chatbots for weight loss, population-based eligibility criteria such as age and weight status were not imposed to allow a discussion on the different needs of an AI chatbot tailored for populations with different demographics. Articles were excluded if they (1) used chatbots that did not incorporate AI (eg, chatbots and computerized coaches that were not conversational and without machine learning capabilities), (2) used human health coaches conversing with users through messaging platforms, (3) did not focus on weight loss or weight loss–related behavior change (eg, diet and exercise), and (4) were on virtual reality or simulation-based conversations and not real-life coaching. The search process and outcomes are shown in Figure 1.

Figure 1. Flow diagram of the search strategy and search outcomes. AI: artificial intelligence.



Stage 4: Charting the Data

Data extraction using Microsoft Excel was first pilot-tested for 3 studies and revised with additional headings before performing data extraction on all the included studies. The headings were author; year; country; type of publication; study design; participant characteristics; sample size; average age; proportion of male participants; baseline BMI; aims; name of the chatbot; delivery mode; use case; architecture; guiding framework; parameters collected; wearables or IoT; availability in multi-language; strategies used to improve user trust, rapport, or emotional connection with the chatbot; device for which the chatbot was deployed; machine learning algorithm or techniques; duration of weight loss program; outcome evaluation; engagement; acceptability; usability or usefulness; user suggestions; and key findings.

Results

Stage 5: Collating, Summarizing, and Reporting the Results

A total of 20 studies were included in this review, of which 1 study comprised 4 separate studies [29], resulting in 23 studies (representing 2231 participants) evaluated in this review. A summary and detailed description of the study characteristics are shown in Table 1 and Multimedia Appendix 3 [25,26,29-47] and Multimedia Appendix 4 [25,26,29-37,39-47]. The chatbot programs included Wakamola [30-32], WaznApp [33], WeightMentor [25], SWITCHes [34], MobileCoach [35], PathMate2 [36], and Lark Weight Loss Health Coach AI [37].

Most (8/23, 35%) of the studies focused on promoting a healthy diet and exercise, whereas only 4% (1/23) studies focused on a healthy diet, exercise, and stress management (Figure 2 and Multimedia Appendix 4). In all, 11 out of the 14 (79%) planned or trialed experimental studies [26,29-33,35-37,40,43-45] reported program durations that ranged from 1 hour to 12 months [30-32]. Only 1 study mentioned the intention of comparing algorithms to yield accurate behavioral predictions [43]. In total, 12 studies mentioned the use of a behavior change framework to guide AI chatbot development (Multimedia Appendix 4). A total of 3 studies used motivational interviewing [26,43,45]; 2 studies used cognitive behavioral therapy [37,45]; and others used mindfulness-based stress reduction [26], dialectic behavior therapy [41], efficiency model of support [39], and control theory by Carver and Scheier [34]. Moreover, 5 studies [29,33] referenced the use of taxonomy of behavior change techniques, whereas the remaining (9/23, 39%) studies did not specify the use of a structured behavior change framework (ie, briefly mentioned the incorporation of behavior change techniques such as goal setting, problem solving, and self-monitoring). Although the value of using a behavior framework to guide the development of weight loss chatbots remains unclear because of the limited number of publications derived from rigorous experimental studies, it could enhance the comprehensiveness of the developed programs and hence, the effectiveness of chatbots in addressing behavior change processes [48]. None of the studies explained the validation process such as using testing or training set splits or k-fold cross-validation.

Table 1. Summary of study characteristics (N=23).

Characteristics	Studies, n (%)
Country	
Ireland [38]	1 (4)
Italy [39]	1 (4)
Lebanon [33]	1 (4)
Spain [30-32]	3 (13)
Switzerland [29,35,36,40]	7 (30)
Taiwan [34]	1 (4)
The Netherlands [41]	1 (4)
United Kingdom [25,42]	2 (9)
United States [26,37,43-46]	6 (26)
Types of publication	
Conference [25,34,35,38-40,42-44]	9 (39)
Internationally peer-reviewed journal articles [29-33,36,37,41,45,46]	14 (61)
Study designs	
Developmental [25,34,35,38,39,43]	6 (26)
Feasibility or pilot [26,30,31]	3 (13)
N-of-1 longitudinal [29]	1 (4)
Observational [32,37,45]	3 (13)
Position or opinion paper [42,46]	2 (9)
Protocol [33]	1 (4)
Qualitative [29,41]	3 (13)
Randomized controlled trials [36,40,44]	3 (13)
Within-subject experiment [29]	1 (4)
Participant characteristics	
Adults with a high BMI [26,37,41,42,44]	5 (22)
Children and adolescents with a high BMI [35,36,45]	3 (13)
General adults [25,29-33]	9 (39)
General children and adolescents [39,40,43]	3 (13)
NS[a] [34,38,46]	3 (13)
Sample sizes	
1-100 [25,26,29,31,35-37,40,41,43-45]	15 (65)
100-800 [30,32]	2 (9)
NS [33,34,36,38,39,42,46]	6 (26)
Age (years)	
<18 [35-37,40,43]	5 (22)
18-40 [26,29-32]	7 (30)
41-65 [25,37,41,44]	4 (17)
NS [29,33,34,38,39,42,46]	7 (30)
Gender (male; %)	
0 [26,41]	2 (9)
<50 [25,29-32,35,37,43-45]	11 (48)
>50 [37,40,43]	3 (13)

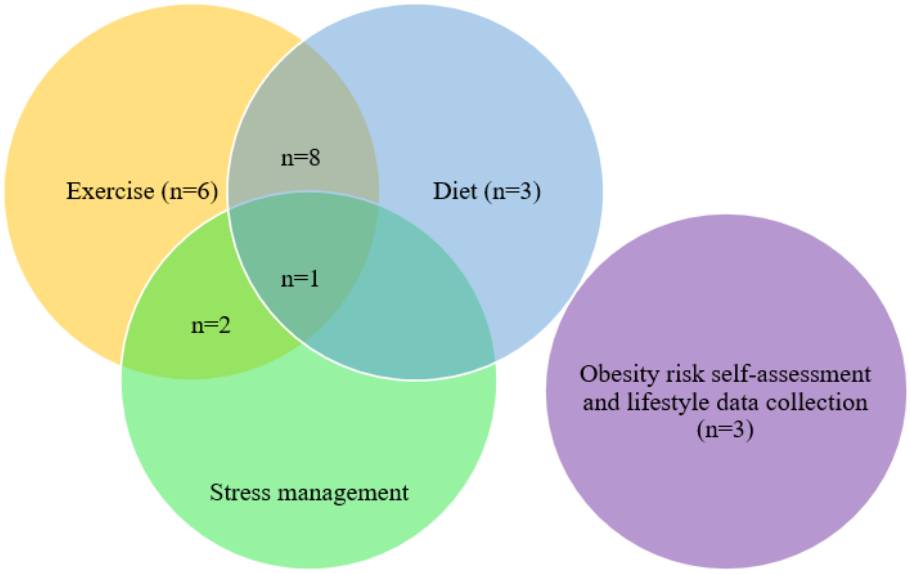
Characteristics	Studies, n (%)
NS [29,33,34,38,39,42,46]	7 (30)
Baseline BMI	
<25 kg/m^2 [31,32]	2 (9)
25-30 kg/m^2 [25,26,30]	3 (13)
>30 kg/m^2 [41,44]	3 (13)
>2 BMI-SDS[b] (remaining studies on children and adolescents did not report BMI) [36,40]	2 (9)
NS [29,33-35,38,39,42,43,45,46]	13 (57)
Mode of delivery	
Speech [43,44]	2 (9)
Text [25,30-33,35-37,39,40,42]	11 (48)
Speech and text [34,45,46]	3 (13)
Text and embodied conversational agent [26]	1 (4)
Speech, text, and AR[c]-embodied conversational agent [29]	4 (17)
NS [38,41]	2 (9)
Multi-language	
Yes [30-32,34]	4 (17)
NS [25,26,29,33,35-46]	19 (83)
Incorporation of wearables or Internet of Things	
Yes [29,33,37,38,40,43]	9 (39)
NS [25,26,30-32,34-36,39,41,42,44-46]	14 (61)
Device used to operationalize chatbots	
Humanoid robot [43]	1 (4)
Smartphone [25,29-31,33,34,36,40,42]	12 (52)
Web browser [26]	1 (4)
NS [32,35,37,38,41,44-46]	9 (39)
Mention of machine learning techniques	
Yes [34,42,43,45]	4 (17)
NS [25,26,29-33,35-41,44,46]	19 (83)
Mention of behavior change framework	
Motivational interviewing [26,43,45]	3 (13)
Cognitive behavioral therapy [37,45]	2 (9)
Mindfulness-based stress reduction [26]	1 (4)
Dialectic behavior therapy [41]	1 (4)
Efficiency model of support [39]	1 (4)
Control theory by Carver and Scheier [34]	1 (4)
Taxonomy of behavior change techniques [29,33]	5 (22)

[a]NS: nonspecified.

[b]SDS: SD score.

[c]AR: augmented reality.

Figure 2. Summary of chatbot use cases for weight loss.



Functions and Architecture

Core functions of the chatbot were to provide personalized weight loss recommendations (20/23, 87%) [25,26,29-34,37-39,41-46] and motivational messages (18/23, 78%) [25,26,29-33,37-39,41-43,45,46] (Multimedia Appendix 5 [25,26,29-37,39-47]). Only 26% (6/23) studies [30-32,36,40,43] mentioned the use of gamification to enhance user engagement, and 26% (6/23) studies [37,39,40,42,43,45] mentioned the use of sentiment analysis to provide emotional support through emotionally appropriate messages (Multimedia Appendix 5). These functions were generally achieved by (1) collecting various user-centric data through chatbot-based self-reports or device-detected metrics, (2) integrating collected parameters using machine learning techniques (including NLP to convert chatbot-collected information for prediction modeling) to predict and generate weight-related recommendations, (3) profiling users according to needs and preferences, and (4) providing personalized chatbot-delivered recommendations. The parameters collected included sociodemographic profiles (eg, age, race, ethnicity, education, work status, and income), food consumption, physical activity (eg, intensity, duration, frequency, and type), stress level, sleep (ie, duration), and clinical profiles (eg, presence of specific chronic diseases, medication use, smoking status, heart rate, and blood pressure; Multimedia Appendix 6 [25,26,29-37,39-47]). Majority of the parameters were collected through chatbots, except in 3 studies that estimated food consumption using smart refrigerator appliances [30] and nutritional information provided by retailers (scanning bar codes) [38,42]; 5 that estimated physical exercise type, intensity, and frequency wearable or smartphone sensors [33,37,38,42,43]; 2 that estimated stress levels using plasma cortisol and skin conductance response [36] and phone detection [37]; and 3 that measured heart rate and blood pressure [36,38,42]. Others that did not mention the use of wearables may have relied on information from in-built sensors of the phone. Only 4 studies elaborated on the algorithms and machine learning techniques used [25,39,42,45].

Outcome Evaluations

Only 4 studies evaluated the effectiveness of a chatbot-delivered program on diet [26,37], physical activity [44], and weight loss [36] (Multimedia Appendix 3). Although 3 of these studies showed greater effectiveness in chatbot-delivered weight loss programs on the measured outcomes, 1 study reported that a higher proportion of adolescents in the control group lost weight as compared with those who interacted with the PathMate 2 chatbot (92% vs 61%). Those in the control group underwent 7 in-person counseling sessions with a health care professional, whereas those in the intervention group interacted with the PathMate 2 chatbot daily with 4 in-person counseling sessions (61%) [36]. Other studies (19/23, 83%) either used chatbots mainly to collect data on diet, physical activity, sitting time, and sleep [31,32] or were still in the developmental stage. Future studies should consider adopting an experimental design that evaluates the use of chatbots on objective weight-related outcomes such as weight loss, diet (eg, food choices, calorie intake, and consumption frequency), and physical activity (eg, energy expenditure, activity type, and activity frequency) using inferential statistics that suggest repeatability. Studies could also explore the mediation and/or moderation effects of these factors including user engagement and satisfaction on weight loss and weight loss maintenance as outcomes to examine the underlying mechanism by which AI chatbots influence weight loss.

Engagement, Satisfaction, and Human-Chatbot Rapport

A total of 6 studies reported estimations of chatbot engagement that averaged at approximately 12 minutes a day [26,45], ranging from 4 minutes to 73 minutes per session [26,30,37,45]. The average daily app use was approximately 71% [36], with more than 4 conversational turns per day [40]. Measures of satisfaction in using the chatbots were heterogeneous, with estimates in terms of willingness to use [43], usability (eg, using the system usability scale) [30,35], adherence to recommendations [26,29,35,36], satisfaction (eg, 4 questions including a net promotor score) [37], and usefulness [45]. A

summary of this section is provided in Multimedia Appendix 5.

An essential element of increasing chatbot engagement was the ability of the chatbot to form a human-chatbot rapport through interactivity and empathy (Multimedia Appendix 7 [25,26,29-37,39-47]). These required the system's capacity to perform sentiment analysis for the interpretation and simulation of culturally appropriate human-like expression of verbal and nonverbal cues (for chatbots with embodiments, eg, speech intonations, facial expressions, and body language). Some techniques used were the deployment of humanoid robots [43] and embodied chatbots [26] that were capable of displaying visual social cues such as eye contact and hand gestures. Other techniques used in text-based chatbots were the delivery of colloquial, personally and culturally appropriate conversational tones and content [26,30]; emojis to emulate human emotional expressions [30-32]; positively framed words [32]; citations of credible information sources [26,33]; and validation (eg, acknowledgments and compliments) of not only behaviors but also thoughts and feelings [41]. Participants of the included studies were also found to have appreciated the personification of the chatbot (eg, funny, animated, empathetic, or playful) [25,29,31,32,37] and the provision of real-time, fast, and reliable recommendations [26,29,38]. In addition, studies (9/23, 39%) that enabled speech, instead of just text-based chatbot interactions (including those that use augmented reality [AR]) [29], improved engagement through a more convenient hands-free voice interaction with the chatbot [34,43-46]. This enabling of conversations through multiple platforms (eg, SMS text messaging, Slack, Telegram, Signal, WhatsApp, or Facebook Messenger) [45] and devices (eg, laptops, Google Home, and Amazon Alexa) [25] has also been reported to increase chatbot engagement because of greater convenience and access.

In contrast, users mentioned concerns regarding privacy and accountability [43,46], the inconvenience of having the chatbot on a limited number of third-party platforms (eg, only Telegram that one may not use) information [31], message or question overload that causes user fatigue [25,31], transparency about the app objectives and information sources [31], and appearing too robotic (eg, speaking too slowly in a robotic voice). Users also suggested that the chatbots should probe further to explore emotions and action plans instead of prescribing them [41]. Most strikingly, users suggested the incorporation of progress-based recommendations, rewards for goals achieved (ie, gamification), and integration with other tracking devices and appliances through IoT [25,26].

Discussion

Principal Findings

Overall, there is a strong potential in AI chatbot–delivered weight loss programs, but more studies are needed to assert sufficient evidence for its implementation in a population that is overweight and obese. The programs captured in this study were heterogeneous in their weight loss use cases, functions, architecture, mode of delivery, and interoperability with other devices and databases. This highlights the need for further research on the impact of various chatbot features such as gamification, personification, and the ability to express empathy and to design and develop an efficient system for weight loss. Most (6/23, 26%) of the studies were still in the development phase (including feasibility testing and qualitative studies on needs and perceptions), with only 3 randomized controlled trials that only reported favorable outcomes of the chatbot on interim user engagement [40], increasing physical activity [44], and weight loss [36]. Only 35% (8/23) of the studies focused on participants with overweight and obesity, 4% (1/23) of the studies were conducted in an Asian context, and most of the studies had a small sample size (15/23, 65%). These gaps raise questions on the receptibility, applicability, and effectiveness of AI chatbots in weight-related behavior change and weight loss in populations with different demographics such as age, weight status, and culture. Among the included studies, 1 study (1/23, 4%) reported that a higher proportion of adolescents in the control group who underwent 7 in-person counseling sessions lost weight as compared with the intervention group who interacted daily with the PathMate 2 chatbot [36]. This finding was contrary to the other 3 studies that reported better diet and exercise improvements in adults who interacted with a chatbot [26,37,44]. Assuming that the improvements in diet and exercise were extrapolated to an eventual weight loss that was not evaluated in the 3 studies, this discrepancy could be associated with adolescents having a lower self-regulation capacity than adults, indicating that chatbot designs must be age appropriate [49]. Having a lower self-regulation capacity may suggest that one requires more frequent and in-person support for impulse control (eg, succumbing to dietary temptations) rather than communicating with a chatbot that is easy to ignore when one is unmotivated. Therefore, chatbot designs for children and adolescents may require more attention-grabbing features such as having an animated embodied CA, more interactivity (ie, engaging as many of the 5 senses as possible) possibly through AR, and gamification to sustain program engagement [50].

Most studies highlighted the use of chatbots to provide personalized nutrition and exercise recommendations and motivational messages, but few studies mentioned the use of gamification and sentiment analysis. Weight loss mobile health apps such as My Fitness Pal and Lifesum are often embellished with gamification features to improve motivation, user engagement, and program effectiveness toward health behavior changes. A study on the 50 most downloaded health apps on the App Store reported that 64% of such apps included some form of goal setting, social presence, challenge, monetary, and social (eg, accomplishing challenges and gaining points to reach higher competition grading tiers) incentives [51]. However, studies have shown that such gamification features do not result in significantly different amounts of weight loss at 3, 6, 9, or 12 months between adults who do and do not undergo such programs [52,53]. Similarly, a meta-analysis reported that gamification did not result in significant weight loss differences between children and adolescents who did and did not undergo gamification for weight loss, although those in the former group were found to have improved nutritional knowledge scores [54]. This suggests that although gamification may improve weight loss knowledge, user engagement, and intention toward health

behavior change, it is insufficient to impact any actual weight loss. Therefore, future studies should focus on identifying more practical and core reasons for weight loss failure, such as the inability to control food temptations, and capitalize on AI chatbot technology to provide real-time nudges.

Major challenges in app-delivered weight loss programs, especially for people with overweight and obesity, lie in users' motivation and discipline toward a diet and exercise regime [55]. Personalization recommendations are well known to enhance goal attainment; hence, mobile health apps strive to provide recommendations based on one's demographic profile, anthropometric status, and monitored calorie intake and output [56]. However, recent studies have shown that this level of personalization is insufficient to sustain weight loss behavior change and that some form of emotional support is required [56]. This is because of the common weight loss–related experiences of stigmatization, self-loathing, and social shaming, which evoke negative emotions such as guilt, shame, self-reproach, regret, depression, anxiety, low self-esteem, and stress [47,57,58]. Such negative emotions could also create a vicious cycle of increasing weight gain, as one copes with such negative emotions by seeking comfort in food. Consistently, poor emotional regulation has been associated with weight regain and weight loss failure, regardless of age, despite the use of behavioral regulation strategies [59,60]. However, current clinical and commercial weight loss programs often neglect this aspect of weight loss, possibly because of the more complex and time-consuming nature. Therefore, interventions that provide emotional support such as health coaching could improve weight loss by forging a supportive coach-client relationship that provides on-demand emotional and knowledge support through accountability, compassion, and empathy [61,62]. However, health coaches are resource intensive and burden the health care system, and the use of AI has been shown to reduce health care costs by increasing health care service delivery efficiencies [63]. Therefore, AI chatbots could supplement the function of health coaches at a lower annualized health care expenditure (eg, through more accurate weight predictions and recommendations, reduced man-hours and infrastructure needed, and reduced admissions). However, this requires chatbots to have enhanced abilities to track emotions through sentiment analysis and emotional modeling to provide empathetic, context-specific messages to motivate health behavior changes, especially in vulnerable situations (eg, in the circumstance of food temptation) [64]. Only 6 of the included studies mentioned the use of sentiment analysis to provide more human-like conversations that consider emotions, and further research is needed to evaluate its effectiveness in improving user engagement and weight loss. The included studies also highlighted some innovative features used to enhance the likeliness of human-like verbal and nonverbal cues such as providing culturally appropriate conversation content; incorporating interactivity and relatability through animations and personified embodied chatbots; and conveying emotions through emojis and body gestures. More research is also needed to evaluate the effects of AI chatbot delivery mode, namely, text-based, speech-based (eg, Alexa), visual (animated 2D characters), and AR-based (animated 3D characters) CAs on user engagement, behavior change, and weight loss.

Practical Recommendations

Overall, chatbots can be programmed to (1) fetch information (eg, weight status, food consumption, and exercise) through conversations with users (eg, asking about food consumed and exercises performed), multiple databases (eg, electronic medical records), devices (eg, activity trackers and smartphones), and smart appliances (eg, smart refrigerators and motion sensors); (2) integrate such information to optimize predictive models of weight loss; (3) synthesize personalized weight loss plans; and (4) provide real-time adaptive recommendations (eg, decision-making and self-regulation skills training), progress feedback (eg, how much more exercise to do to reach a certain weight loss goal by a stipulated time), and emotional support (eg, motivation, empowerment, and validation) through conversations with users.

Limitations

Certain relevant evidence could have been precluded from this study, undermining the comprehensiveness of this review, although many databases including gray literature were searched. This includes programs that were commercialized and marketed without a research study and studies published in other languages. Studies included in this review were also largely heterogeneous in study design, participant characteristics, and outcomes measured, impeding the comparisons between AI chatbot elements, weight-related outcome measures, and architectures to inform future chatbot designs and developments. However, this also highlights the infancy and potential of such technology in reducing the health care burden of overweight and obesity, a long-standing public health problem.

Conclusions

This study highlighted the potential of AI chatbots in providing just-in-time personalized weight loss–related behavior change recommendations, motivational messages, and emotional support. These require the integration of a comprehensive set of information beyond the conventional health metrics from self-reports, app trackers, and fitness wearables. This includes personality and preferences (eg, based on goal achievements), circumstantial behaviors (eg, trigger-based overconsumption), and emotional states (eg, chatbot conversations and wearable stress detectors). AI chatbots should be designed to be human-like, personalized, contextualized, immersive, and enjoyable to enhance user experience, engagement, behavior change, and weight loss. Future AI chatbot developments should also consider issues of privacy; accountability; user burden during chatbot engagement; and interoperability with other databases (eg, electronic medical records), third-party apps (eg, health tracking apps), social media platforms (eg, data mining from Twitter, Facebook, and Instagram posts), devices (eg, laptops, desktops, and phones), and appliances (eg, refrigerators and gaming consoles). Future AI chatbots should also be designed as a one-stop diet, exercise, and emotional support app to derive at a market-ready and effective chatbot-delivered weight loss program.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 23 KB - medinform_v10i4e32578_app1.docx]

Multimedia Appendix 2
Details on search terms used for each database.
[DOCX File , 24 KB - medinform_v10i4e32578_app2.docx]

Multimedia Appendix 3
Detailed characteristics of each study (N=23).
[DOCX File , 30 KB - medinform_v10i4e32578_app3.docx]

Multimedia Appendix 4
Use cases, names, delivery modes, deployment devices, wearables or Internet of Things and behavioral framework of the conversational agents used in the included studies (N=23).
[DOCX File , 24 KB - medinform_v10i4e32578_app4.docx]

Multimedia Appendix 5
Architecture or descriptions and core functions of conversational agents used in the included studies (N=23).
[DOCX File , 28 KB - medinform_v10i4e32578_app5.docx]

Multimedia Appendix 6
Parameters collected from users to develop conversational agent-based weight loss interventions (N=23).
[DOCX File , 25 KB - medinform_v10i4e32578_app6.docx]

Multimedia Appendix 7
Engagement, satisfaction and human-conversational agent emotional connection described in the included studies (N=23).
[DOCX File , 26 KB - medinform_v10i4e32578_app7.docx]

References

1. Blüher M. Obesity: global epidemiology and pathogenesis. Nat Rev Endocrinol 2019 May;15(5):288-298. [doi: 10.1038/s41574-019-0176-8] [Medline: 30814686]
2. Obesity and overweight. World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2022-04-04]
3. Norris T, Cole TJ, Bann D, Hamer M, Hardy R, Li L, et al. Duration of obesity exposure between ages 10 and 40 years and its relationship with cardiometabolic disease risk factors: a cohort study. PLoS Med 2020 Dec;17(12):e1003387 [FREE Full text] [doi: 10.1371/journal.pmed.1003387] [Medline: 33290405]
4. Wearing SC, Hennig EM, Byrne NM, Steele JR, Hills AP. Musculoskeletal disorders associated with obesity: a biomechanical perspective. Obes Rev 2006 Aug;7(3):239-250. [doi: 10.1111/j.1467-789X.2006.00251.x] [Medline: 16866972]
5. Argolo DF, Hudis CA, Iyengar NM. The impact of obesity on breast cancer. Curr Oncol Rep 2018 Apr 11;20(6):47. [doi: 10.1007/s11912-018-0688-8] [Medline: 29644507]
6. Dietz W, Santos-Burgoa C. Obesity and its implications for COVID-19 mortality. Obesity (Silver Spring) 2020 Jun;28(6):1005. [doi: 10.1002/oby.22818] [Medline: 32237206]
7. Kroes M, Osei-Assibey G, Baker-Searle R, Huang J. Impact of weight change on quality of life in adults with overweight/obesity in the United States: a systematic review. Curr Med Res Opin 2016;32(3):485-508. [doi: 10.1185/03007995.2015.1128403] [Medline: 26652030]
8. Pereira-Miranda E, Costa PR, Queiroz VA, Pereira-Santos M, Santana ML. Overweight and obesity associated with higher depression prevalence in adults: a systematic review and meta-analysis. J Am Coll Nutr 2017;36(3):223-233. [doi: 10.1080/07315724.2016.1261053] [Medline: 28394727]
9. Simmonds M, Llewellyn A, Owen CG, Woolacott N. Predicting adult obesity from childhood obesity: a systematic review and meta-analysis. Obes Rev 2016 Feb;17(2):95-107. [doi: 10.1111/obr.12334] [Medline: 26696565]

10. Chew HS, Lopez V. Global impact of COVID-19 on weight and weight-related behaviors in the adult population: a scoping review. Int J Environ Res Public Health 2021 Feb 15;18(4):1876 [FREE Full text] [doi: 10.3390/ijerph18041876] [Medline: 33671943]
11. Sherman RP, Petersen R, Guarino AJ, Crocker JB. Primary care-based health coaching intervention for weight loss in overweight/obese adults: a 2-year experience. Am J Lifestyle Med 2017 Jun 19;13(4):405-413 [FREE Full text] [doi: 10.1177/1559827617715218] [Medline: 31285724]
12. Mirkarimi K, Kabir MJ, Honarvar MR, Ozouni-Davaji RB, Eri M. Effect of motivational interviewing on weight efficacy lifestyle among women with overweight and obesity: a randomized controlled trial. Iran J Med Sci 2017 Mar;42(2):187-193 [FREE Full text] [Medline: 28360445]
13. Vallabhan MK, Jimenez EY, Nash JL, Gonzales-Pacheco D, Coakley KE, Noe SR, et al. Motivational interviewing to treat adolescents with obesity: a meta-analysis. Pediatrics 2018 Nov;142(5):e20180733 [FREE Full text] [doi: 10.1542/peds.2018-0733] [Medline: 30348753]
14. Barnes RD, Ivezaj V, Martino S, Pittman BP, Paris M, Grilo CM. Examining motivational interviewing plus nutrition psychoeducation for weight loss in primary care. J Psychosom Res 2018 Jan;104:101-107 [FREE Full text] [doi: 10.1016/j.jpsychores.2017.11.013] [Medline: 29275778]
15. Barnes RD, Ivezaj V, Martino S, Pittman BP, Grilo CM. Back to basics? No weight loss from motivational interviewing compared to nutrition psychoeducation at one-year follow-up. Obesity (Silver Spring) 2017 Dec;25(12):2074-2078 [FREE Full text] [doi: 10.1002/oby.21972] [Medline: 29086484]
16. Patel ML, Wakayama LN, Bass MB, Breland JY. Motivational interviewing in eHealth and telehealth interventions for weight loss: a systematic review. Prev Med 2019 Sep;126:105738. [doi: 10.1016/j.ypmed.2019.05.026] [Medline: 31153917]
17. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
18. Chew HS, Ang WH, Lau Y. The potential of artificial intelligence in enhancing adult weight loss: a scoping review. Public Health Nutr 2021 Jun;24(8):1993-2020 [FREE Full text] [doi: 10.1017/S1368980021000598] [Medline: 33592164]
19. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv (forthcoming) 2018.
20. Budzianowski P, Vulić I. Hello, it's GPT-2--how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. arXiv (forthcoming) 2019.
21. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Mach (Dordr) 2020 Nov 01;30(4):681-694. [doi: 10.1007/s11023-020-09548-1]
22. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems 32. 2019 Presented at: NeurIPS '19; December 8-14, 2019; Vancouver, Canada.
23. Turing-NLG: a 17-billion-parameter language model by Microsoft. Microsoft. 2020. URL: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/ [accessed 2022-04-04]
24. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
25. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, McTear M. IEEE International Conference on Bioinformatics and Biomedicine. 2019 Presented at: BIBM '19; November 18-21, 2019; San Diego, CA, USA p. 2845-2851. [doi: 10.1109/BIBM47256.2019.8983073]
26. Gardiner PM, McCue KD, Negash LM, Cheng T, White LF, Yinusa-Nyahkoon L, et al. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Educ Couns 2017 Sep;100(9):1720-1729 [FREE Full text] [doi: 10.1016/j.pec.2017.04.015] [Medline: 28495391]
27. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
28. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
29. Kowatsch T, Lohse KM, Erb V, Schittenhelm L, Galliker H, Lehner R, et al. Hybrid ubiquitous coaching with a novel combination of mobile and holographic conversational agents targeting adherence to home exercises: four design and evaluation studies. J Med Internet Res 2021 Feb 22;23(2):e23612 [FREE Full text] [doi: 10.2196/23612] [Medline: 33461957]
30. Asensio-Cuesta S, Blanes-Selva V, Conejero A, Portolés M, García-Gómez M. A user-centered chatbot to identify and interconnect individual, social and environmental risk factors related to overweight and obesity. Inform Health Soc Care 2022 Jan 02;47(1):38-52. [doi: 10.1080/17538157.2021.1923501] [Medline: 34032537]
31. Asensio-Cuesta S, Blanes-Selva V, Conejero JA, Frigola A, Portolés MG, Merino-Torres JF, et al. A user-centered chatbot (Wakamola) to collect linked data in population networks to support studies of overweight and obesity causes: design and pilot study. JMIR Med Inform 2021 Apr 14;9(4):e17503 [FREE Full text] [doi: 10.2196/17503] [Medline: 33851934]

32. Asensio-Cuesta S, Blanes-Selva V, Portolés M, Conejero JA, García-Gómez JM. How the Wakamola chatbot studied a university community's lifestyle during the COVID-19 confinement. Health Informatics J 2021;27(2):14604582211017944 [FREE Full text] [doi: 10.1177/14604582211017944] [Medline: 34044657]
33. Bardus M, Hamadeh G, Hayek B, Al Kherfan R. A self-directed mobile intervention (WaznApp) to promote weight control among employees at a Lebanese university: protocol for a feasibility pilot randomized controlled trial. JMIR Res Protoc 2018 May 16;7(5):e133 [FREE Full text] [doi: 10.2196/resprot.9793] [Medline: 29769174]
34. Huang CY, Yang MC, Huang CY, Chen YJ, Wu ML, Chen KW. A chatbot-supported smart wireless interactive healthcare system for weight control and health promotion. In: IEEE International Conference on Industrial Engineering and Engineering Management. 2018 Presented at: IEEM '18; December 16-19, 2018; Bangkok, Thailand p. 1791-1795. [doi: 10.1109/IEEM.2018.8607399]
35. Kowatsch T, Volland D, Shih I, Rüegger D, Künzler F, Barata F, et al. Design and evaluation of a mobile chat app for the open source behavioral health intervention platform mobilecoach. In: Bertino E, Gao W, Steffen B, Woeginger G, Yung M, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Berlin, Germany: Springer; 2017:485-489.
36. Stasinaki A, Büchter D, Shih CH, Heldt K, Güsewell S, Brogle B, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 09;21(1):308 [FREE Full text] [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]
37. Stein N, Brooks K. A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. JMIR Diabetes 2017 Nov 01;2(2):e28 [FREE Full text] [doi: 10.2196/diabetes.8590] [Medline: 30291087]
38. Wu Y, Donovan R, Vu B, Engel F, Hemmje M, Afli H. Chatbot based behaviour analysis for obesity support platform. In: Proceedings of the 6th Collaborative European Research Conference. 2020 Presented at: CERC '20; September 10-11, 2020; Belfast, UK p. 112-124.
39. Fadhil A, Gabrielli S. Addressing challenges in promoting healthy lifestyles: the AI-chatbot approach. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2017 Presented at: PervasiveHealth '17; May 23-26, 2017; Barcelona, Spain p. 261-265. [doi: 10.1145/3154862.3154914]
40. L'Allemand D, Shih CH, Heldt K, Buchter D, Brogle B, Rüegger D, et al. Design and interim evaluation of a smartphone app for overweight adolescents using a behavioural health intervention platform. Obes Rev 2018;19(Suppl 1):102-107.
41. Dol A, Bode C, Velthuijsen H, van Strien T, van Gemert-Pijnen L. Application of three different coaching strategies through a virtual coach for people with emotional eating: a vignette study. J Eat Disord 2021 Jan 14;9(1):13 [FREE Full text] [doi: 10.1186/s40337-020-00367-4] [Medline: 33446275]
42. Sandri S, Zheng H, Engel F, Moorhead A, Wang H, Bond R, et al. Is there an optimal technology to provide personal supportive feedback in prevention of obesity? In: IEEE International Conference on Bioinformatics and Biomedicine. 2019 Presented at: BIBM '19; November 18-21, 2019; San Diego, CA, USA p. 1-6. [doi: 10.1109/BIBM47256.2019.8983405]
43. Addo ID, Ahamed SI, Chu WC. Toward collective intelligence for fighting obesity. In: IEEE 37th Annual Computer Software and Applications Conference. 2013 Presented at: COMPSAC '13; July 22-26, 2013; Kyoto, Japan p. 690-695. [doi: 10.1109/COMPSAC.2013.109]
44. Hassoon A, Baig Y, Naimann D, Celentano D, Lansey D, Stearns V, et al. Abstract 54: addressing cardiovascular health using artificial intelligence: randomized clinical trial to increase physical activity in cancer survivors using intelligent voice assist (Amazon Alexa) for patient coaching. Circulation 2020 Mar 03;141(Suppl_1):A54. [doi: 10.1161/circ.141.suppl_1.54]
45. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019 May 16;9(3):440-447. [doi: 10.1093/tbm/ibz043] [Medline: 31094445]
46. Thompson D, Baranowski T. Transl Behav Med 2019 May 16;9(3):448-450. [doi: 10.1093/tbm/ibz065] [Medline: 31094432]
47. Wu YK, Berry DC, Schwartz TA. Weight stigmatization and binge eating in Asian Americans with overweight and obesity. Int J Environ Res Public Health 2020 Jun 17;17(12):4319 [FREE Full text] [doi: 10.3390/ijerph17124319] [Medline: 32560329]
48. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37 [FREE Full text] [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
49. Blakemore SJ, Choudhury S. Development of the adolescent brain: implications for executive function and social cognition. J Child Psychol Psychiatry 2006;47(3-4):296-312. [doi: 10.1111/j.1469-7610.2006.01611.x] [Medline: 16492261]
50. Del Río NG, González-González CS, Martín-González R, Navarro-Adelantado V, Toledo-Delgado P, García-Peñalvo F. Effects of a gamified educational program in the nutrition of children with obesity. J Med Syst 2019 May 22;43(7):198. [doi: 10.1007/s10916-019-1293-6] [Medline: 31119385]
51. Cotton V, Patel MS. Gamification use and design in popular health and fitness mobile applications. Am J Health Promot 2019 Mar;33(3):448-451 [FREE Full text] [doi: 10.1177/0890117118790394] [Medline: 30049225]

52. Patel MS, Small DS, Harrison JD, Hilbert V, Fortunato MP, Oon AL, et al. Effect of behaviorally designed gamification with social incentives on lifestyle modification among adults with uncontrolled diabetes: a randomized clinical trial. JAMA Netw Open 2021 May 03;4(5):e2110255 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.10255] [Medline: 34028550]
53. Kurtzman GW, Day SC, Small DS, Lynch M, Zhu J, Wang W, et al. Social incentives and gamification to promote weight loss: the LOSE IT randomized, controlled trial. J Gen Intern Med 2018 Oct;33(10):1669-1675 [FREE Full text] [doi: 10.1007/s11606-018-4552-1] [Medline: 30003481]
54. Suleiman-Martos N, García-Lara RA, Martos-Cabrera MB, Albendín-García L, Romero-Béjar JL, Cañadas-De la Fuente GA, et al. Gamification for the improvement of diet, nutritional habits, and body composition in children and adolescents: a systematic review and meta-analysis. Nutrients 2021 Jul 20;13(7):2478 [FREE Full text] [doi: 10.3390/nu13072478] [Medline: 34371989]
55. Edney S, Ryan JC, Olds T, Monroe C, Fraysse F, Vandelanotte C, et al. User engagement and attrition in an app-based physical activity intervention: secondary analysis of a randomized controlled trial. J Med Internet Res 2019 Nov 27;21(11):e14645 [FREE Full text] [doi: 10.2196/14645] [Medline: 31774402]
56. Asimakopoulos S, Asimakopoulos G, Spillers F. Motivation and user engagement in fitness tracking: heuristics for mobile healthcare wearables. Informatics 2017 Jan 22;4(1):5. [doi: 10.3390/informatics4010005]
57. Barbarin AM, Saslow LR, Ackerman MS, Veinot TC. Toward health information technology that supports overweight/obese women in addressing emotion- and stress-related eating. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, Canada p. 1-14. [doi: 10.1145/3173574.3173895]
58. Papadopoulos S, Brennan L. Correlates of weight stigma in adults with overweight and obesity: a systematic literature review. Obesity (Silver Spring) 2015 Sep;23(9):1743-1760 [FREE Full text] [doi: 10.1002/oby.21187] [Medline: 26260279]
59. Reinelt T, Petermann F, Bauer F, Bauer CP. Emotion regulation strategies predict weight loss during an inpatient obesity treatment for adolescents. Obes Sci Pract 2020 Jun;6(3):293-299 [FREE Full text] [doi: 10.1002/osp4.410] [Medline: 32523718]
60. Sainsbury K, Evans EH, Pedersen S, Marques MM, Teixeira PJ, Lähteenmäki L, et al. Attribution of weight regain to emotional reasons amongst European adults with overweight and obesity who regained weight following a weight loss attempt. Eat Weight Disord 2019 Apr;24(2):351-361 [FREE Full text] [doi: 10.1007/s40519-018-0487-0] [Medline: 29453590]
61. McQueen A, Imming ML, Thompson T, Garg R, Poor T, Kreuter MW. Client perspectives on health coaching: insight for improved program design. Am J Health Behav 2020 Sep 01;44(5):591-602. [doi: 10.5993/AJHB.44.5.4] [Medline: 33121578]
62. Sieczkowska SM, de Lima AP, Swinton PA, Dolan E, Roschel H, Gualano B. Health coaching strategies for weight loss: a systematic review and meta-analysis. Adv Nutr 2021 Jul 30;12(4):1449-1460. [doi: 10.1093/advances/nmaa159] [Medline: 33339042]
63. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
64. Kim J, Oh U. EmoWei: emotion-oriented personalized weight management system based on sentiment analysis. In: IEEE 20th International Conference on Information Reuse and Integration for Data Science. 2019 Presented at: IRI '19; July 30-August 1, 2019; Los Angeles, CA, USA p. 342-349. [doi: 10.1109/IRI.2019.00060]

Abbreviations

AI: artificial intelligence
AR: augmented reality
CA: conversational agent
IoT: Internet of Things
NLP: natural language processing
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

Edited by C Lovis; submitted 02.08.21; peer-reviewed by Y Li, N Maglaveras, J Poon; comments to author 28.09.21; revised version received 04.10.21; accepted 08.01.22; published 13.04.22.

Please cite as:
Chew HSJ
The Use of Artificial Intelligence–Based Conversational Agents (Chatbots) for Weight Loss: Scoping Review and Practical Recommendations
JMIR Med Inform 2022;10(4):e32578
URL: https://medinform.jmir.org/2022/4/e32578
doi:10.2196/32578
PMID:35416791

Review

Machine Learning Approach for Preterm Birth Prediction Using Health Records: Systematic Review

Zahra Sharifi-Heris[1], MSN; Juho Laitala[2], MSc; Antti Airola[2], PhD; Amir M Rahmani[1], PhD; Miriam Bender[1], PhD

[1]Sue & Bill Gross School of Nursing, University of California, Irvine, CA, United States

[2]Department of Computing, University of Turku, Turku, Finland

Corresponding Author:
Zahra Sharifi-Heris, MSN
Sue & Bill Gross School of Nursing
University of California
802 W Peltason Dr
Irvine, CA, 92697
United States
Phone: 1 6506805432
Email: sharifiz@uci.edu

Abstract

Background: Preterm birth (PTB), a common pregnancy complication, is responsible for 35% of the 3.1 million pregnancy-related deaths each year and significantly affects around 15 million children annually worldwide. Conventional approaches to predict PTB lack reliable predictive power, leaving >50% of cases undetected. Recently, machine learning (ML) models have shown potential as an appropriate complementary approach for PTB prediction using health records (HRs).

Objective: This study aimed to systematically review the literature concerned with PTB prediction using HR data and the ML approach.

Methods: This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. A comprehensive search was performed in 7 bibliographic databases until May 15, 2021. The quality of the studies was assessed, and descriptive information, including descriptive characteristics of the data, ML modeling processes, and model performance, was extracted and reported.

Results: A total of 732 papers were screened through title and abstract. Of these 732 studies, 23 (3.1%) were screened by full text, resulting in 13 (1.8%) papers that met the inclusion criteria. The sample size varied from a minimum value of 274 to a maximum of 1,400,000. The time length for which data were extracted varied from 1 to 11 years, and the oldest and newest data were related to 1988 and 2018, respectively. Population, data set, and ML models' characteristics were assessed, and the performance of the model was often reported based on metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve.

Conclusions: Various ML models used for different HR data indicated potential for PTB prediction. However, evaluation metrics, software and package used, data size and type, selected features, and importantly data management method often remain unjustified, threatening the reliability, performance, and internal or external validity of the model. To understand the usefulness of ML in covering the existing gap, future studies are also suggested to compare it with a conventional method on the same data set.

(JMIR Med Inform 2022;10(4):e33875) doi:10.2196/33875

KEYWORDS

preterm birth; prediction model; machine learning approach; artificial intelligence

Introduction

Background

Preterm birth (PTB), a common pregnancy complication, is responsible for 1.085 million (35%) of the 3.1 million neonatal deaths each year and significantly affects approximately 15 million children annually worldwide [1]. Survivors often suffer from lifetime disabilities, including motor function problems, learning disabilities, and visual and hearing dysfunctions [2]. In almost all high- and middle-income countries, PTB and its

adverse consequences are the major leading causes of death in children aged <5 years [2]. According to the World Health Organization, PTB is defined as birth before 37 completed weeks of gestation (<259 days) from the first day of a woman's last menstrual period. In general, there is a negative association between gestational age and poor pregnancy outcomes and long-term complications such as hospitalization, longer stay in the neonatal intensive care unit, and death [2]. Long-term hospitalization and frequent medical services required for PTB survivors may lead to additional mental distress and extra costs for the family, and it also imposes more strain on the health care system [3]. Current screening tests for PTB prediction can be categorized into three main groups: (1) risk factor evaluation, (2) cervical measurement, and (3) biochemical biomarker assessment. However, not all approaches have potential to be translated into clinical predictive utility, safely and cost-effectively [4]. They may also be insufficient for detecting true-positive PTB cases. For example, biochemical assessment is a costly procedure that may impose physical and mental stress to the pregnant individual. Risk factor assessment is another commonly used approach for which information comes from evidence-based practice that is an end outcome of statistical hypothesis testing (often including 1 factor to be tested) under controlled settings, which is a time- and money wasting approach. The latter may also leave behind many potential risk factors that did not receive researchers' attention, advancing to hypothesis testing. By contrast, previous PTB history is one of the dominant risk factors, with a relative risk of 13.56, leaving nulliparous women undetected [3,5]. These findings indicate the insufficiency of the current methods in predicting high-risk pregnancies, specifically in those who are experiencing their first pregnancy. A few predictive systems have also been studied using series of information including maternal demographics, medical and obstetrical history, and well-known risk factors; unfortunately, however, their predictive power has been very limited [6,7]. This limitation may be because they often rely on simple linear statistical models that lack the capacity to model complex problems such as PTB. It is suggested that risk factor assessment using conventional approaches is insufficient, as >50% of PTB pregnancies will fail to be identified [8]. Thus, identifying additional screening tools for covering the gap in conventional prediction approaches is highly critical, as it helps guide prenatal care and prepare for potential early interventions required for poor prognosis. Recently, machine learning (ML) methods have been applied to further improve individual risk prediction beyond traditional models. Many ML methods can model the complex nonlinear relationships between the predictor features and the outcome. ML techniques can learn the structure from data without being explicitly programmed for its function [9]. For the ML approach, a significant volume of data is required to create robust models with high accuracy.

Objectives

Fortunately, health records (HRs) in most countries contain data regarding one's sociodemographic, obstetric, and medical history. This makes HRs appropriate data sets for ML models to learn and eventually predict the intended outcome. There has been growing research on applied ML on HR data to identify efficient predictive models for the early diagnosis of PTB. Few systematic or literature reviews, although are informative, are not focused on PTB [10]. This systematic review article aims to review the literature that has attempted to use ML on HR data to predict mothers who are at risk for PTB.

Methods

Overview

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. A comprehensive search was performed in bibliographic databases including PubMed, CINAHL, MEDLINE, Web of Science, Scopus, Engineering Village (Compendex and Inspec), and IEEE Computer Society Digital Library, until May 15, 2021, in collaboration with a medical librarian (Stephen L Clancy). The search terms included controlled and free-text terms. The search strategy and number of articles found from each database are shown in Multimedia Appendix 1. Two review authors (ZSH and JL) independently performed the title or abstract and full-text screening. Potential disagreements were resolved by a third independent researcher. Nonrelevant articles were excluded in the title and abstract screening, and for the full-text article screen, reasons for exclusion per article were recorded. References of the identified articles were also checked for potential additional papers. Data were extracted by ZSH and confirmed by JL. Discrepancies were revisited by both authors to guarantee the database accuracy.

Eligibility Criteria and Study Selection

Studies were included if they aimed to predict PTB risk by using HR data. The outcome variable was PTB occurrence, which is globally defined as any pregnancy termination between 20 and 37 weeks of gestation. Although in some studies PTB was defined differently in terms of age range, all definitions were aligned under 37 weeks of gestational age. The PTB definition serves to examine and establish model performance (ie, the ability of the intended model to distinguish PTB cases from non-PTB cases). The papers were required to include a statement of the ML domain or any of its synonyms. To identify any study that failed to include a ML statement in the title or abstract, an extensive list of commonly used ML model techniques was added to the search strategy.

Selection Process

Selected articles were peer reviewed in the Covidence web-based software [11] by 2 independent reviewers. To assess relevancy, all studies were screened based on titles, abstracts, and full texts in two steps. In the first step, the abstracts of all articles gathered from the databases were screened in terms of their relevance to our study aim. Next, those articles with relevant titles or abstracts resulting from the first step underwent a full-text assessment. To resolve the raised disagreement, a third reviewer was involved for consulting. All articles that were concerned with heart rate variability assessment during pregnancy were included.

Quality of Evidence

The quality of studies was assessed using the criteria proposed by Qiao [12]. Although the criteria proposed by Qiao [12] were too restrictive, no other quality assessment tool was found for the quality assessment of the studies. In this approach, quality assessment is based on five different categories: unmet needs, reproducibility, robustness, generalizability, and clinical significance. Unmet needs are met if the limits were reported in current non-ML approaches (eg, current methods have low diagnostic accuracy). A study is considered reproducible if it describes used feature engineering methods, platforms and packages, and hyperparameters. The condition for robustness is fulfilled if valid methods are used to overcome the overfitting (*k*-fold cross-validation or bootstrap*:* when a data set is large, splitting it into separate training, validation, and test sets is the best approach [13], and *k*-fold cross-validation and bootstrap are required only with the small data sets when there are not enough data for a 3-way split [14]) and the stability of results (variation of the validation statistic) are reported. The generalizability condition is met if the model is validated using external data. A study is considered to have clinical significance if predictors are explained and clinical applications for the model are suggested. Quality assessment was conducted by providing a *yes* or *no* response for each of the 5 categories. However, in our study, we attempted to be more descriptive; thus, a short description was provided for some of the criteria when applicable in the quality assessment table (Table 1).

Table 1. Quality assessment.

Study	Unmet need (existing gap)	Reproducibility			Robustness		Generalizability (external validation data)	Clinical significance	
		Feature engineering	Platform package	Hyperparameters	Valid methods to overcome overfitting	Stability of results		Predictor explanation	Suggested clinical use
Weber et al, 2018 [15]	Yes	Yes	Yes	No	5-fold CV[a]	Minimum and maximum values reported from the CV	No	Logistic regression coefficients and odds ratios	No
Rawashdeh et al, 2020 [16]	Yes	Yes	Yes	Number of neighbors for KNN[b], number of hidden layers for ANN[c], number of trees for RF[d]	Train-test split. Train size 237 with 19 positives. Test size 37 with 7 positives	No	No	No	Yes
Gao et al, 2019 [17]	Yes	Representing medical concepts as a bag of words and word embeddings, TF-IDF[e], discretization of continuous features	No	No	Train-test split. Train size 17,607 with 132 positives. Test size 8082 with 85 positives	Minimum and maximum values and CIs	No	Feature importance, odds ratio	Yes
Lee and Ahn, 2019 [18]	Yes	No	Yes	Only neural network architecture described	Train-test split. Both train and test sets contained 298 participants	No	No	Feature importance (RF and ANN)	No
Woolery and Grzymala-Busse, 1994 [19]	Yes	No	Yes	No	A total of 3 different data sets used in isolation; 50-50 train-test split was used with each data set	No	No	No	No
Grzymala-Busse and Woolery, 1994 [20]	Yes	No	Yes	No	A total of 3 different data sets used in isolation; 50-50 train-test split was used with each data set	No	No	No	No

Study	Unmet need (existing gap)	Reproducibility			Robustness		Generalizability (external validation data)	Clinical significance	
		Feature engineering	Platform package	Hyperparameters	Valid methods to overcome overfitting	Stability of results		Predictor explanation	Suggested clinical use
Vovsha et al, 2014 [21]	Yes	No	Yes	No	Data separated timewise to 3 data sets, and 80-20 train-test split was used with each data set; 5-fold CV to select models	No	No	Feature importance (linear SVM[f])	No
Esty et al, 2018 [22]	Yes	No	Yes	No	No	No	No	No	No
Frize et al, 2011 [23]	Yes	No	Yes	No	Division into 3 data sets (parous and nulliparous). Train-test-verification splits	SDs of the metrics were reported	No	No	No
Goodwin and Maher, 2000 [24]	Yes	No	Yes	No	Train-test split (75%-25%)	No	No	Feature importance	No
Tran et al, 2016 [3]	Yes	Unigrams were created from free-text fields after removal of stop words	No	No	Train-test split (66%-33%)	No	No	Feature importance	Yes
Koivu and Sairanen, 2020 [9]	Yes	New features were created. Continuous features were standardized, and nominal features were one-hot encoded	Yes	All hyperparameters described	Data set partitioned into 4 parts (feature selection, training, validation, and test, with stratified splits of 10%-70%-10%-10%)	95% CIs for metrics	Yes	Feature importance	Yes
Khatibi et al, 2019 [25]	Yes	Imputation with mode for categorical features and median for continuous features	No	No	Train-test split	No	No	Feature importance	No

[a]CV: cross-validation.

[b]KNN: *K*-nearest neighbor.

[c]ANN: artificial neural network.

[d]RF: random forest.

[e]TF-IDF: term frequency-inverse document frequency.

[f]SVM: support vector machine.

Data Synthesis

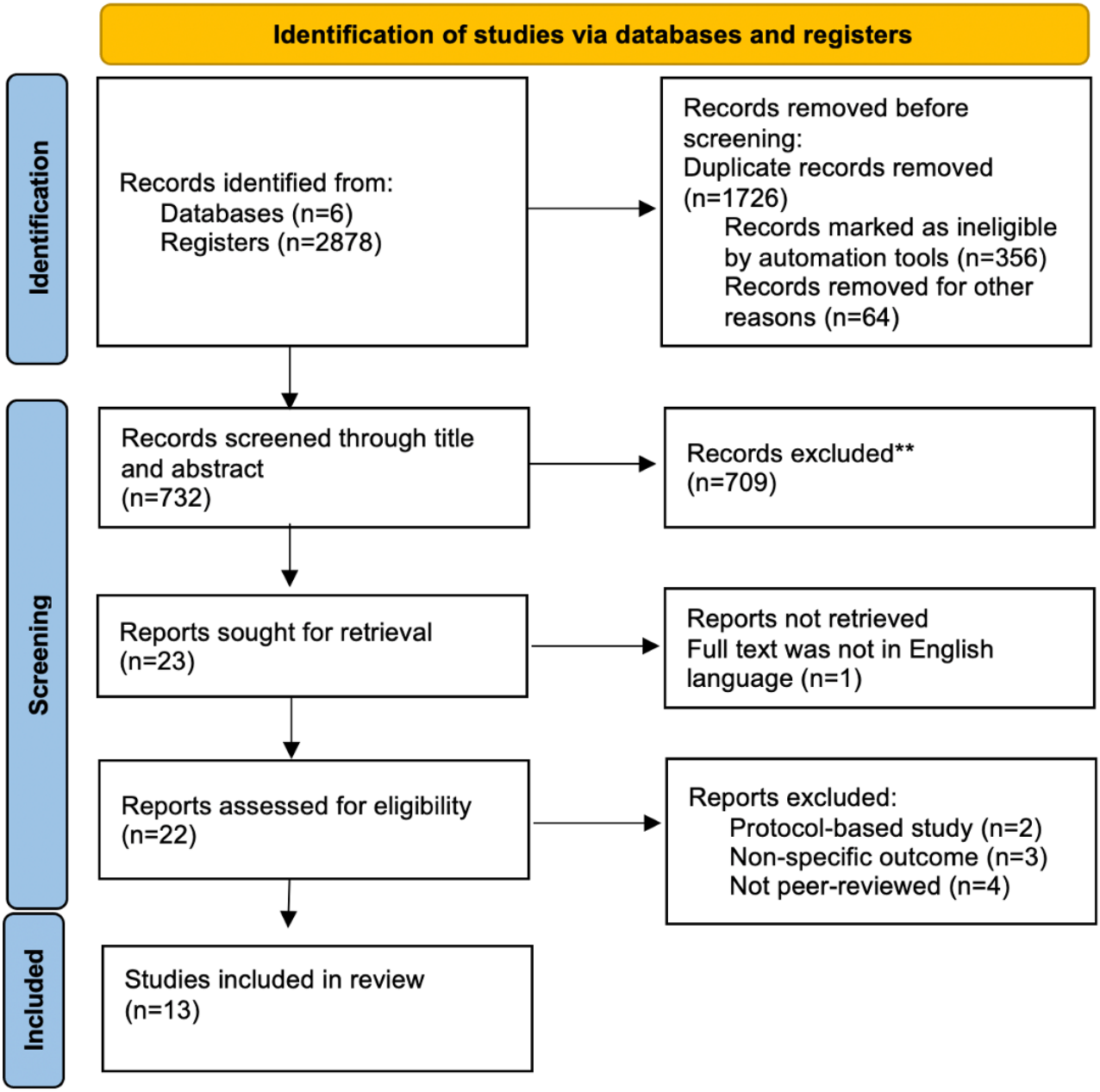
The reviewed studies were not homogenous in terms of methodology and data set; thus, a meta-analysis was not possible. A narrative synthesis was chosen to bring together broad knowledge from various approaches. This type of synthesis is not the same as a narrative description that accompanies many reviews. To synthesize the literature, we applied a guideline from Popay et al [26]. The steps included (1) preliminary analysis, (2) exploration of relationships, and (3) assessment of the robustness of the synthesis. Theory development was not performed because of the exploratory nature of the research synthesized. Thematic analysis was applied to extract the main themes from all the studies. The two main themes developed in the results represent the main areas of knowledge available regarding ML models applied for PTB prediction during pregnancy. These included *descriptive characteristics of the data set* (eg, data source, population, case and control definition, and feature selection) and *ML methodologies* (eg, feature selection, model processing, performance evaluation, and findings). We could not compare the studies because of the divergence of studies in terms of data set, ML model processing, and evaluation metric. The quality of the papers was assessed using the method proposed by Qiao [12].

Results

Study Selection

After removing duplicates, 732 papers were screened through title and abstract. Of these 732 studies, 23 (3.1%) were screened by full text, resulting in 13 (1.8%) papers that met the inclusion criteria. Reasons for exclusion at this stage were recorded and are shown in the flow diagram in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) chart.



Study Characteristics

All the studies were retrospective and used one or more data sets recorded in clinical settings. Of the 13 studies, 7 (54%) were conducted in or after 2018 and 9 (69%) originated from the United States. The time length for which data were extracted varied from 1 to 11 years, and the oldest and newest data were related to 1988 and 2018, respectively. Of the 13 studies, 6 (46%) did not report the ethnicity or race of the population whose data were modeled. Various data sets were used for the studies, and the number of data sets varied from 1 to 3 in each study. The types of information included in each data set varied, including demographic, obstetric history, medical background, and clinical and laboratory information. Demographic information was included in almost all of the data sets used in the included studies. The size of the population whose data have been used for ML modeling varied from 274 to 13,150,017 people, and the number of features considered for modeling varied from 19 to 5000 depending on the data set used. PTB was defined differently from study to study; the cutoff point for the control and study groups (PTB and non-PTB) was defined as the 37th week of gestational age for 77% (10/13) of the studies that matched the standard cutoff point between term and PTBs. Of the 13 studies, 3 (23%) determined the PTB cutoff based on the frequency of the newborn death [17], newborn viability chance [16], or no justification [20]. It was not always specified whether abortion (pregnancy termination <20 weeks) was included in the models. Indeed, there was often no clear discernment of abortion and PTB in the reviewed studies (see Table 2 for more details).

Table 2. Descriptive characteristics of studies and feature selection.

Study, country, and type of study	Population characteristics	Data source (number of features)	Population (birth)	Study (PTB[a]), control groups, and type of PTB	Feature selection process and gestational week for when selected features are related	Number of selected features	Date
Weber et al, 2018 [15], United States, retrospective	Nulliparous women with a singleton birth (<32, ≥20, and ≥37 weeks); non-Hispanic Black (n=54,084) and White (n=282,130)	Birth certificate and hospital discharge records: >1000 features	336,214	PTB (early spontaneous): ≥20 and <32 weeks; control: ≥37 weeks	Factors with uncertain and ambiguous values were excluded, highly correlated features were collapsed, exclusion of features with no variation; —[b]	20	2007 to 2011
Rawashdeh et al, 2020 [16], Australia, retrospective	Australian; pregnancies with cervical cerclage	Data from a fetal medicine unit in a tertiary hospital in NSW[c]: 19 features	274	PTB (spontaneous): <26 weeks; control: >26 weeks	Unnecessary features (eg, medical record numbers) were excluded	19	2003 to 2014
Gao et al, 2019 [17], United States, retrospective	Caucasian (>68%), Black (16%-21%), and other (10%-13%)	EHR[d] of Vanderbilt University Medical Center: 150 features	25,689	PTB: <28 weeks; control: ≥28 weeks; type of PTB was not distinguished	Features were arranged by their information gain and top 150 features were retained; —	150	2005 to 2017
Lee and Ahn, 2019 [18], Korea, retrospective	Korean; induced labors were excluded	Anam Hospital in Seoul	596	PTB (spontaneous): >20 and <37 weeks; control: ≥37 weeks	—	14	2014 to 2018
Woolery and Grzymala-Busse, 1994 [19], United States, retrospective	—	3 data sets: 214 features in total	18,890	PTB: <37 weeks; control: ≥37 weeks; type of PTB was not distinguished	—	Data set 1 (n=52), data set 2 (n=77), and data set 3 (n=85)	1994
Grzymala-Busse and Woolery, 1994 [20], United States, retrospective	—	3 data sets:153 features in total	9480	PTB: <36 weeks; control: ≥36 weeks; type of PTB was not distinguished	—	Data set 1 (n=13), data set 2 (n=73), and data set 3 (n=67)	1994
Vovsha et al, 2014 [21], United States, retrospective	—	NICHD[e]-MFMU[f] data set: >400 features	2929	PTB (spontaneous and induced): <32, <35, and <37 weeks; control: ≥37 weeks	Logistic regression with forward selection, stepwise selection, LASSO[g], and elastic net; —	24th week (n=50), 26th week (n=205), and 28th week (n=316)	1992 to 1994
Esty et al, 2018 [22], United States and Canada, retrospective	—	BORN[h] and PRAMS[i]: 520 features	782,000	PTB: <37 weeks; control: ≥37 weeks; type of PTB was not distinguished	Features with >50% missing values were removed before missing value imputation; features come from before the 23rd gestational week	520	—
Frize et al, 2011 [23], United States, retrospective	—	PRAMS: >300 features	>113, 000	PTB: <37 weeks; control: ≥37 weeks; type of PTB was not distinguished	Decision tree (to establish consistency between data sets, features specific to the United States were excluded, eg, Medicaid and Women Infants Children Program); features come from before the 23rd gestational week	19 for parous and 16 for nulliparous	2002 to 2004

Study, country, and type of study	Population characteristics	Data source (number of features)	Population (birth)	Study (PTB[a]), control groups, and type of PTB	Feature selection process and gestational week for when selected features are related	Number of selected features	Date
Goodwin and Maher, 2000 [24], United States, retrospective	—	Duke University's Medical Center TMR TM perinatal data: 4000−5000 features	63,167	PTB: <37 weeks; control: ≥37 weeks; type of PTB was not distinguished	Heuristic techniques (features related to week <37 were included); —	32 demographic and 393 clinical	1988 to 1997
Tran et al, 2016 [3], Australia, retrospective	Australian	RNS[j], NSW	15,814 births	PTB (spontaneous and elective): <34 and <37 weeks; control: ≥37 weeks	Features kept based on their importance (top *k* features; [27]); the rare features that occur in <1% of data points were removed; features come from before the 25th gestational week	10	2011 to 2015
Koivu and Sairanen, 2020 [9], United States, retrospective	White, Black, American Indian or Alaskan native, and Asian or Pacific Island individuals	CDC[k] and NYC[l] data sets	13,150,017	PTB: <37 weeks; control: ≥37 weeks; type of PTB was not distinguished	Excluding highly correlated features with correlation analysis (Pearson); —	26	CDC: 2013 to 2016; NYC: 2014 to 2016
Khatibi et al 2019 [25], Iran, retrospective	Iranian	National maternal and neonatal records (IMaN[m] registry): 112 features	>1,400,000	PTB (spontaneous and medically indicated): >28 and <37 weeks; control: ≥37 weeks	Parallel feature selection and classification methods including MR_PB-PFS (features with nonzero scores are selected as top features); —	112	2016 to 2017

[a]PTB: preterm birth.

[b]Not reported in the study.

[c]NSW: New South Wales.

[d]EHR: electronic health record.

[e]NICHD: National Institute of Child Health and Human Development.

[f]MFMU: Maternal-Fetal Medicine Units Network.

[g]LASSO: least absolute shrinkage and selection operator.

[h]BORN: Better Outcomes Registry Network.

[i]PRAMS: Pregnancy Risk Monitoring Assessment System.

[j]RNS: Royal North Shore.

[k]CDC: Centers for Disease Control and Prevention.

[l]NYC: New York City.

[m]IMaN: Iranian Maternal and Neonatal Network.

Data Selection

Of the 13 studies, 9 (69%) reported at least one piece of preprocessing information regarding the included data. The preprocessing step included data mapping, missing data management, and the class imbalance management in data. For the feature selection, of the 13 studies, 11 (85%) reported at least one method for the feature selection process. The number of features selected for each study varied from 10 to 520 for final ML modeling. On the basis of the literature surveyed, of the 13 studies, only 2 (15%) used unsupervised feature selection. In addition, of the 13 studies, 3 (23%) did not use feature selection, and some studies did use some heuristics instead. Owing to the divergency in feature selection, we could not identify clear trends on how the used approach would affect the model performance (see Table 3 for more information).

Table 3. Data processing and machine learning modeling.

Study	Preprocessing data		Model	Dominant model	Evaluation metrics	Analysis software and package	Findings
	Missing data management	Class imbalance					
Weber et al, 2018 [15]	MICE[a]	—[b]	Super learning approach using logistic regression, random forest, *K*-nearest neighbors, LR[c] (LASSO[d], ridge, and an elastic net)	No difference between models	Sensitivity, specificity, PVP[e], PVN[f], and AUC[g]	Rstudio (version 3.3.2), SuperLearner package	AUC=0.67, sensitivity=0.61, specificity=0.64
Rawashdeh et al, 2020 [16]	Instances with missing values were removed manually	SMOTE[h]	Locally weighted learning, Gaussian process, K-star classifier, linear regression, *K*-nearest neighbor, decision tree, random forest, neural network	Random forest	Accuracy, sensitivity, specificity, AUC, and G-means	WEKA[i] (version 3.9)	Random forest: G-mean=0.96, sensitivity=1.00, specificity=0.94, accuracy=0.95, AUC=0.98 (oversampling ratio of 200%)
Gao et al, 2019 [17]	—	Control group were undersampled	RNNs[j], long short-term memory network, logistic regression, SVM[k], Gradient boosting	RNN ensembled models on balanced data	Sensitivity, specificity, PVP, and AUC	—	AUC=0.827, sensitivity=0.965, specificity=0.698, PVP=0.033
Lee and Ahn, 2019 [18]	—	—	ANN[l], logistic regression, decision tree, naïve Bayes, random forest, SVM	No difference between models	Accuracy	Python (version 3.52)	No difference in accuracy between ANN (0.9115) with logistic regression and the random forest (0.9180 and 0.8918, respectively)
Woolery and Grzymala-Busse, 1994 [19]	—	—	LERS[m]	—	Accuracy	ID3[n], LERS CONCLUS	Database 1: accuracy=88.8% accurate for both low-risk and high-risk pregnancy. Database 2: accuracy=59.2% in high-risk pregnant women. Database 3: accuracy=53.4%
Grzymala-Busse and Woolery,1994 [20]	—	—	LERS based on the *bucket brigade algorithm* of genetic algorithms and enhanced by partial matching	—	Accuracy	LERS	Accuracy=68% to 90%
Vovsha et al, 2014 [21]	—	Oversampling techniques (Adasyn)	SVMs with linear and nonlinear kernels, LR (forward selection, stepwise selection, L1 LASSO regression, and elastic net regression)	—	Sensitivity, specificity, and G-means	Rstudio, glmnet package	SVM: sensitivity (0.404 to 0.594), specificity (0.621 to 0.84), G-mean (0.575 to 0.652); LR: sensitivity (0.502 to 0.591), specificity (0.587 to 0.731), G-mean (0.586 to 0.604)

Study	Preprocessing data		Model	Dominant model	Evaluation metrics	Analysis software and package	Findings
	Missing data management	Class imbalance					
Esty et al, 2018 [22]	Imputation with the *missForest* package in R	Not clear	Hybrid C5.0 decision tree−ANN classifier	—	Sensitivity, specificity, and ROC[o]	R software, missForest Package, FANN[p] library	Sensitivity: 84.1% to 93.4%, specificity: 70.6% to 76.9%, AUC: 78.5% to 89.4%
Frize et al, 2011 [23]	Decision tree	—	Hybrid decision tree–ANN	—	Sensitivity, specificity, ROC for P[q] and NP[r] cases	See5, MATLAB Neural Ware tool	Training (P: sensitivity=66%, specificity=83%, AUC=0.81; NP: sensitivity=62.8%, specificity=71.7%, AUC=0.72), test (P: sensitivity=66.3%, specificity=83.9%, AUC=0.80; NP: sensitivity=65%, specificity=71.3%, AUC=0.73), and verification (P sensitivity=61.4%, specificity=83.3%, AUC=0.79; NP: sensitivity=65.5%, specificity=71.1%, AUC=0.73)
Goodwin and Maher, 2000 [24]	PVRuleMinerl or FactMiner	—	Neural networks, LR, CART[s], and software programs called PVRuleMiner and FactMiner	No difference between models	ROC	Custom data mining software (Clinical Miner and PVRuleMiner, FactMiner)	No significant difference between techniques. Neural network (AUC=0.68), stepwise LR (AUC=0.66), CART (AUC=0.65), FactMiner (demographic features only; AUC=0.725), FactMiner (demographic plus other indicator features; AUC=0.757)
Tran et al, 2016 [3]	—	Undersampling of the majority class	SSLR[t], RGB[u]	—	Sensitivity, specificity, NPV[v], PVP, F-measure, and AUC	—	SSLR: sensitivity=0.698 to 0.734, specificity=0.643 to 0.732, F-measure=0.70 0.73, AUC=0.764 to 0.791, NPV=0.96 to 0.719, PVP=0.679, 0.731; RGB: sensitivity=0.621 to 0.720, specificity=0.74 to 0.841, F-measures=0.693 to 0.732, NPV=0.675 to 0.717, PVP=0.783 to 0.743, AUC=0.782 to 0.807

Study	Preprocessing data		Model	Dominant model	Evaluation metrics	Analysis software and package	Findings
	Missing data management	Class imbalance					
Koivu and Sairanen, 2020 [9]	—	—	LR, ANN, LGBM[w], deep neural network, SELU[x] network, average ensemble, and weighted average WA[y] ensemble	—	AUC	Rstudio (version 3.5.1) and Python (version 3.6.9)	AUC for classifiers: LR=0.62 to 0.64; deep neural network: 0.63 to 0.66; SELU network: 0.64 to 0.67; LGBM: 0.64 to 0.67; average ensemble: 0.63 to 0.67; WA ensemble: 0.63 to 0.67
Khatibi et al, 2019 [25]	Map phase module	—	Decision trees, SVMs and random forests, ensemble classifiers	—	Accuracy and AUC	—	Accuracy=81% and AUC=68%

[a]MICE: Multiple Imputation by Chained Equations.
[b]Not reported in the study.
[c]LR: linear regression.
[d]LASSO: least absolute shrinkage and selection operator.
[e]PVP: predictive value positive.
[f]PVN: predictive value negative.
[g]AUC: area under the ROC curve.
[h]SMOTE: Synthetic Minority Oversampling Technique.
[i]WEKA: Waikato Environment for Knowledge Analysis.
[j]RNN: recurrent neural network.
[k]SVM: support vector machine.
[l]ANN: artificial neural network.
[m]LERS: learning from examples of rough sets.
[n]ID3: iterative dichotomiser 3.
[o]ROC: receiver operating characteristic.
[p]FANN: Fast Artificial Neural Network.
[q]P: parous.
[r]NP: nulliparous.
[s]CART: classification and regression tree.
[t]SSLR: stabilized sparse logistic regression.
[u]RGB: Randomized Gradient Boosting.
[v]NPV: net present value.
[w]LGBM: Light Gradient Boosting Machine.
[x]SELU: scaled exponential linear unit.
[y]WA: weighted average.

Identified Potential Risk Factors

Although the included features somewhat differed in the studies, some features were commonly used and considered potential risk factors that may predict PTB occurrence (Table 4).

Table 4. Frequency of potential risk factors in the studies (n=13).

Potential risk factors	Studies, n (%)
Previous PTB[a]	10 (77)
Hypertensive disorders	9 (70)
Maternal age	7 (54)
Cervical or uterus disorders (cerclage, myoma, or inconsistency)	7 (54)
Ethnicity and race	6 (46)
Diabetes (eg, gestational, mellitus)	6 (46)
Smoking or substance abuse	5 (38)
Multiple pregnancy	5 (38)
Education	4 (30)
Physical characteristics (BMI, weight, and height)	4 (30)
Parity	4 (30)
Marital status	3 (23)
Other chronic diseases (thyroid, asthma, systemic lupus erythematosus, or cardiovascular)	3 (23)
PTB symptoms (bleeding, contractions, premature rupture of membranes, etc)	3 (23)
Insurance	2 (15)
Income	2 (15)
In vitro fertilization	2 (15)
Stress or domestic violence	2 (15)
Infections (gonorrhea, syphilis, chlamydia, or hepatitis C)	1 (7)
Biopsy	1 (7)

[a]PTB: preterm birth.

ML Modeling and Performance Assessment

Various basic and complex ML modeling approaches were used with different frequencies, including artificial neural network, logistic regression, decision tree, support vector machine (SVM) with linear and nonlinear kernels, linear regression (least absolute shrinkage and selection operator [LASSO], ridge, and elastic net), random forest, locally weighted learning, gradient boosting, learning from examples of rough sets, Gaussian process, K-star classifier, and naïve Bayes (Multimedia Appendix 2).

Although most studies reported the type of software applied for the ML analysis, only few of them specified the package they have used for the analysis. Several evaluation measures were used to assess the proposed models. These include sensitivity, specificity, area under the receiver operating characteristic curve, accuracy, predictive value positive, predictive value negative, G-mean, F-measure, and net present value, based on the frequency they have been used in the studies. Owing to the divergent methodology used for outcome assessment and model processing, comparison between models was not possible. However, overall, studies with a cutoff gestational age of 37th week, regardless of the model used, often showed lower sensitivity (40%-69%), except for 1 study that showed a sensitivity of 93% [22]. Those with an earlier cutoff gestational age of 26th to 28th weeks indicated higher sensitivity (96%-100%).

Quality Assessment

In general, reviewed studies had satisfactory quality (Table 1). However, there was substantial variation, as some studies fulfilled almost every category, whereas others met only a few. All studies fulfilled *the unmet need category*, as PTB prediction is still an unsolved problem. Feature engineering was mentioned in almost half (6/13, 46%) of the studies [3,9,15-17,25]. Platforms and packages were not mentioned in 23% (3/13) of the studies [3,17,25]. Hyperparameters were described in only 23% (3/13) of the studies [9,16,18]. According to the criteria proposed by Qiao [12], of the 13 studies, only 1 (8%) used valid methods (*k*-fold cross-validation) to overcome overfitting [15]. However, many of the studies have population sizes of tens of thousands or higher, which makes the standard train-test split a valid approach for model evaluation, and there was no need for *k*-fold cross-validation. There is no commonly agreed criterion for sufficiency of data for a single train-test split to be sufficient, as this depends on factors such as number of features, relative sizes of the classes, and amount of noise in the data. As an example, previously, Kohavi [28] studied the accuracy estimation and model selection with the test set size of 500 instances as the lower limit for a single train-test split being considered reliable. In 23% (3/13) of the studies, the use of *k*-fold cross-validation or bootstrap instead of the train-test split would have been clearly the better choice because of the small population size (n<3000) [16,18,21]. The stability of the results is reported only for 31% (4/13) of the studies [9,15,17,23]. Of

the 13 studies, only 1 (8%) used external validation data and met the requirement for generalizability [9]. Predictor explanation was provided in 62% (8/13) of the studies [3,9,15,17,18,21,24,25]. Only 31% (4/13) of the studies clearly suggested a clinical application for their method [3,9,16,17].

Discussion

Principal Findings

Premature birth remains a public health concern worldwide. Survivors experience substantial lifetime morbidity and mortality rates. The conventional methods of PTB assessment that have been used by clinicians seem to be insufficient to identify PTB risk in more than half of the cases. The conventional methods that are concerned with health data (HR) are often statistical modeling, in which, first, input predictive factors are selected by a researcher and, second, the multifactorial nature of PTB is ignored. Thus, these methods suffer from biases and linearities. The linear vision on HR in conventional approaches is perhaps one of the major barriers to advancing our understanding of nonlinear interaction dynamics between potential risk factors of multifactorial PTB. ML modeling, in contrast to statistical modeling, investigates the structure of the target phenomenon without preassumption on data, and automatically and thoroughly explores possible nonlinear associations and higher-order interactions (more than 2-way) between potential the risk factors and the outcome [29]. ML modeling is expected to discover novel patterns, not necessarily novel predictive features, which provide an opportunity to gain insight into the underlying mechanisms of multifactorial outcomes (in this case PTB), where existing knowledge is still insufficient for developing a thorough predictive system [29]. Over the past 26 years, 13 studies have been published, creating ML-based prediction models using HR data, with the number of studies increasing over time.

Among the reviewed studies, the performance of various ML modeling indicated potential for predictive purposes. Owing to the different evaluation metrics used by studies, performance comparison across studies was not practical. On the basis of within-study synthesis, some studies compared nonlinear ML methods, such as deep neural networks, kernel SVMs, or random forests, to more basic linear models, such as logistic regression, LASSO, and elastic net. Of these 13 studies, 4 (31%) concluded that there was no significant difference between the predictive performances of the different applied methods [3,9,19,21]. For example, Tran et al [3] compared stabilized sparse logistic regression with randomized gradient boosting and found no significant differences between the methods. The conclusion that complex ML modeling is not superior to simple logistic modeling matches the findings of a recent systematic review conducted for a wider concept of clinical prediction. In the aforementioned review, Christodolou et al [30] compared the performance of logistic regression with more complex ML-based clinical prediction models; they found no evidence of the superior performance of the ML methods for clinical prediction. In contrast, some studies indicated a significant difference among various ML modeling approaches. For example, Rawashdeh et al [16] showed that random forest has a clear advantage over linear regression in predicting the week of delivery; however, the test set used in the study was very small for a reliable conclusion. Vovsha et al [21] also showed some improvements for nonlinear SVM over a linear model (linear SVM, LASSO, and elastic net) when classifying preterm versus full-term birth for the whole study population but did not find similar differences when making predictions for only spontaneous PTB or for first-time mothers. Gao et al [17] and Koivu and Sairanen [9] reported that deep learning–based approaches have better performance than logistic regression. The remaining studies did not include a comparison with a basic baseline method, such as logistic regression. In conclusion, these results imply that classical statistical models remain a competitive approach for predicting PTB. The current limitations of ML modeling and its infancy may explain its failure to cover the gaps in classical statistical models for PTB prediction using HR data. We suggest that more research is still required to ascertain with confidence whether ML methods, such as those based on deep learning, can systematically improve the predictive performance of the model as compared with basic statistical models.

An HR seems to be a useful data source, including the potential risk factors from which the ML model can learn the significant predictors as well as the nonlinear interaction among the identified risk factors.

A large sample size, as one of the distinct characteristics of HR data, is a double-edged sword that covers large populations but consumes time and requires advanced technology. A large data size can also be used to create validation sets. Most studies in this review had large sample sizes, including thousands of pregnant women. Although some studies performed internal validation, external validation was uncommon, and almost all studies validated the performance within the same HR. Th lack of external validity assessment limits generalizability and may reduce the discrimination validity of the model when applied in other sites and HR systems. External validation of the model through its application in a distinct data set may be helpful in understanding its usefulness and generalizability in different geographical areas, periods, and settings [31]. Furthermore, half of the studies in this review did not report the race or ethnicity of the population, which indicates ignoring the importance of the ethnic and health disparity in predictive model assessment. For example, ethnic minority groups, such as Black and Hispanic women, are more at risk of developing pregnancy complications, including PTB. Failure to consider ethnicity threatens the internal validity of ML modeling.

Large data sizes and reflective data types are as important as large sample sizes. HR data often appear insufficient to precisely identify risk factors that decrease the accuracy of predictive ML models. Indeed, small sample size and passive data that are limited to a few sociodemographic and medical histories seem insufficient to predict the multifactorial PTB. Enriched data that include more, time-sensitive, and dynamic characteristics of each individual (eg, life history, mental distress during various stages of pregnancy, and biomarker change) may increase the accuracy and integrity of the applied ML models. For example, being diagnosed with gestational diabetes is known to be a strong predictive factor for PTB among the features in ML

models. However, owing to the dynamic nature of diabetes (glucose level), which can vary from moment to moment, particularly during pregnancy, applying a pool of data reflecting the dynamic glucose change in a person may be more accurate in predicting PTB in comparison with the presence or absence of diabetes. The difference in glucose change may also partially explain why some women with diabetes are at a higher risk of developing PTB. To achieve this accuracy in HR use, data should be enriched by more and dynamic features and ML models should be optimized to analyze the dynamic-natured potential risk factors that go beyond the clear-cut presence or absence of a feature [32].

In contrast, a small data size threatens the risk factor distinction for PTB prediction. There might be an indirect association between some predictive factors and PTB, falsifying the direct and actual associations. For example, smoking not only is introduced as a protective factor against mortality in low–birth weight and PTB infants but also is identified as a predictive factor for PTBs. In this case, PTB may not be the result of smoking directly itself but due to potential mediators, such as hypertension, which is triggered by smoking. Therefore, if there is no recorded information about blood pressure, the model may consider smoking as the actual risk factor. This highlights the importance of more possible health data to increase the ability of the ML model to distinguish between mediators and exposure features.

One of the major challenges in HR-based studies is the presence of missing data. Although missing data have been an acknowledged challenge in HR studies, a little more than half of the studies acknowledged the presence of missing data and a variety of analytic approaches to manage this absence. On average, despite its importance, there has been minimal work in this area, and it is unclear how such biased observations impact prediction models.

Another important challenge in HR-related models is unbalanced data between case and control groups. This problem is because PTB occurs in 10% of all births. Researchers have often applied oversampling techniques to handle unbalanced data. However, these techniques create artificial data that may not have much in common with actual observations. Oversampling techniques must be used carefully in validating models because if artificial instances end up in the test set (or test folds in cross-validation), one may obtain highly overoptimistic performance estimates.

In addition, all reviewed studies approached PTB prediction as a classification problem. There was often no clear discernment of abortion and PTB in the reviewed studies. This ambiguity, if it comes from missing to distinguish abortion from PTB in actual ML modeling, may threaten the specificity of the model in predicting PTB. In addition, as PTB and abortion have different leading causes, the findings of the studies may also be questionable. In addition, in the defined PTB time window (20-37 gestational week), classification remains problematic. In this case, neonates born at week ≤30 are considered to belong to the same class as those born at week 36 of pregnancy. However, the former is associated with a much higher risk of adverse outcomes and requires neonatal intensive care. Therefore, it could be more beneficial to approach PTB as a regression problem and try to predict the gestational age (as weeks or days) at childbirth. This approach could help identify PTB cases that have the greatest need for care.

Conclusions

Overall, ML modeling has been indicated to be a potentially useful approach in predicting PTB, although future studies are suggested to minimize the aforementioned limitations to achieve more accurate models. Importantly, ML's ability to cover the existing gap in conventional statistical methods remains questionable. To achieve reliable conclusions, our study suggests some considerations for future studies. First, more studies are needed to compare ML modeling with existing conventional methods in the same data set with the same amount of data and population. Conducting the comparison studies uncovers the potential superiority of one over the other. Second, the study population should be distinguished based on parity, particularly if previous pregnancy data were among the selected features. Otherwise, the model would probably rely on this strong predictive factor in multiparous women, leaving nulliparous women underserved and undetected. In addition, studies should be transparent to whether they use the same time frame for feature selection for case (PTB) and control (non-PTB) groups. For instance, assume that we have a cutoff point of 28 weeks before which we want our model to identify PTB cases. In this case, if we include the data for the control group to be after the cutoff point, which most likely differs from before the cutoff point, the model may rely on the information after the cutoff point for PTB prediction. Thus, the model fails to detect the cases before the specified time point. Third, two cutoff points should be clarified in model development: (1) the gestational cutoff week the study targets before the cases are detected and (2) the gestational time point before the features are selected. For example, Gao et al [17] determined the 28th week as the cutoff week before feature selection. However, it is not clear whether the created model would identify PTB before week 28, from where the features were collected, or any time before week 37, based on the data related to before the 28th week. The time interval between identified features and PTB occurrence, particularly if the PTB is symptomatic, can be more informative in terms of model specificity and time sensitivity in detecting symptomatic and asymptomatic PTB.

Enriched data size and optimized data type can also improve the usefulness of the ML model. Appropriate approaches for managing missing data and unbalanced control and case groups are also required to achieve more reliable and accurate results.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 17 KB - medinform_v10i4e33875_app1.docx]

Multimedia Appendix 2
Machine learning models' frequency.
[PNG File , 615 KB - medinform_v10i4e33875_app2.png]

References

1. Walani SR. Global burden of preterm birth. Int J Gynaecol Obstet 2020 Jul;150(1):31-33. [doi: 10.1002/ijgo.13195] [Medline: 32524596]
2. Preterm birth. World Health Organization. 2018. URL: https://www.who.int/news-room/fact-sheets/detail/preterm-birth [accessed 2022-01-21]
3. Tran T, Luo W, Phung D, Morris J, Rickard K, Venkatesh S. Preterm birth prediction: deriving stable and interpretable rules from high dimensional data. In: Proceedings of the 1st Machine Learning for Healthcare Conference. 2016 Presented at: PMLR '16; August 19-20, 2016; Los Angeles, CA, USA p. 164-177.
4. FIGO Working Group on Good Clinical Practice in Maternal-Fetal Medicine. Good clinical practice advice: prediction of preterm labor and preterm premature rupture of membranes. Int J Gynaecol Obstet 2019 Mar;144(3):340-346. [doi: 10.1002/ijgo.12744] [Medline: 30710365]
5. Esplin MS, O'Brien E, Fraser A, Kerber RA, Clark E, Simonsen SE, et al. Estimating recurrence of spontaneous preterm delivery. Obstet Gynecol 2008 Sep;112(3):516-523. [doi: 10.1097/AOG.0b013e318184181a] [Medline: 18757647]
6. Mercer BM, Goldenberg RL, Das A, Moawad AH, Iams JD, Meis PJ, et al. The preterm prediction study: a clinical risk assessment system. Am J Obstet Gynecol 1996 Jun;174(6):1885-1895. [doi: 10.1016/s0002-9378(96)70225-9] [Medline: 8678155]
7. Lee KA, Chang MH, Park MH, Park H, Ha EH, Park EA, et al. A model for prediction of spontaneous preterm birth in asymptomatic women. J Womens Health (Larchmt) 2011 Dec;20(12):1825-1831 [FREE Full text] [doi: 10.1089/jwh.2011.2729] [Medline: 22023413]
8. Georgiou HM, Di Quinzio MK, Permezel M, Brennecke SP. Predicting preterm labour: current status and future prospects. Dis Markers 2015;2015:435014 [FREE Full text] [doi: 10.1155/2015/435014] [Medline: 26160993]
9. Koivu A, Sairanen M. Predicting risk of stillbirth and preterm pregnancies with machine learning. Health Inf Sci Syst 2020 Dec;8(1):14 [FREE Full text] [doi: 10.1007/s13755-020-00105-9] [Medline: 32226625]
10. Sufriyana H, Husnayain A, Chen YL, Kuo CY, Singh O, Yeh TY, et al. Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: systematic review and meta-analysis. JMIR Med Inform 2020 Nov 17;8(11):e16503 [FREE Full text] [doi: 10.2196/16503] [Medline: 33200995]
11. Better systematic review management. Covidence. URL: https://www.covidence.org/ [accessed 2022-01-21]
12. Qiao N. A systematic review on machine learning in sellar region diseases: quality and reporting items. Endocr Connect 2019 Jul;8(7):952-960 [FREE Full text] [doi: 10.1530/EC-19-0156] [Medline: 31234143]
13. Hastie T, Friedman J, Tibshirani R. The elements of statistical learning. New York, NY, USA: Springer; 2001.
14. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test 2018;2(3):249-262 [FREE Full text] [doi: 10.1007/s41664-018-0068-2] [Medline: 30842888]
15. Weber A, Darmstadt GL, Gruber S, Foeller ME, Carmichael SL, Stevenson DK, et al. Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. Ann Epidemiol 2018 Nov;28(11):783-9.e1. [doi: 10.1016/j.annepidem.2018.08.008] [Medline: 30236415]
16. Rawashdeh H, Awawdeh S, Shannag F, Henawi E, Faris H, Obeid N, et al. Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage. Comput Biol Chem 2020 Apr;85:107233. [doi: 10.1016/j.compbiolchem.2020.107233] [Medline: 32106071]
17. Gao C, Osmundson S, Velez Edwards DR, Jackson GP, Malin BA, Chen Y. Deep learning predicts extreme preterm birth from electronic health records. J Biomed Inform 2019 Dec;100:103334 [FREE Full text] [doi: 10.1016/j.jbi.2019.103334] [Medline: 31678588]
18. Lee KS, Ahn KH. Artificial neural network analysis of spontaneous preterm labor and birth and its major determinants. J Korean Med Sci 2019 Apr 29;34(16):e128 [FREE Full text] [doi: 10.3346/jkms.2019.34.e128] [Medline: 31020816]
19. Woolery LK, Grzymala-Busse J. Machine learning for an expert system to predict preterm birth risk. J Am Med Inform Assoc 1994;1(6):439-446 [FREE Full text] [doi: 10.1136/jamia.1994.95153433] [Medline: 7850569]
20. Grzymala-Busse JW, Woolery LK. Improving prediction of preterm birth using a new classification scheme and rule induction. Proc Annu Symp Comput Appl Med Care 1994:730-734 [FREE Full text] [Medline: 7950021]

21. Vovsha I, Rajan A, Salleb-Aouissi A, Raja A, Radeva A, Diab H, et al. Predicting preterm birth is not elusive: machine learning paves the way to individual wellness. In: Big Data Becomes Personal: Knowledge into Meaning: Papers from the AAAI Spring Symposium. 2014 Presented at: AAAI '14; March 24–26, 2014; Palo Alto, CA, USA p. 82-89.
22. Esty A, Frize M, Gilchrist J, Bariciak E. Applying data preprocessing methods to predict premature birth. In: Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2018 Presented at: EMBC '18; July 18-21, 2019; Honolulu, HI, USA p. 6096-6099. [doi: 10.1109/embc.2018.8513681]
23. Frize M, Yu N, Weyand S. Effectiveness of a hybrid pattern classifier for medical applications. Int J Hybrid Intell Syst 2011 May 04;8(2):71-79. [doi: 10.3233/his-2011-0123]
24. Goodwin L, Maher S. Data mining for preterm birth prediction. In: Proceedings of the 2000 ACM symposium on Applied computing - Volume 1. 2000 Presented at: SAC '00; March 19-21, 2000; Como, Italy p. 46-51. [doi: 10.1145/335603.335680]
25. Khatibi T, Kheyrikoochaksarayee N, Sepehri MM. Analysis of big data for prediction of provider-initiated preterm birth and spontaneous premature deliveries and ranking the predictive features. Arch Gynecol Obstet 2019 Dec;300(6):1565-1582. [doi: 10.1007/s00404-019-05325-3] [Medline: 31650230]
26. Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme. Peninsula Medical School, Universities of Exeter and Plymouth. 2006. URL: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf [accessed 2022-04-04]
27. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat 2008 Sep 1;2(3):916-954. [doi: 10.1214/07-AOAS148]
28. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. 1995 Presented at: IJCAI'95; August 20-25, 1995; Montreal, Canada p. 1137-1143.
29. Ryo M, Rillig MC. Statistically reinforced machine learning for nonlinear patterns and variable interactions. Ecosphere 2017 Nov 27;8(11):e01976. [doi: 10.1002/ecs2.1976]
30. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
31. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015 Jan;68(1):25-34. [doi: 10.1016/j.jclinepi.2014.09.007] [Medline: 25441703]
32. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics 2000 Dec;1(4):465-480. [doi: 10.1093/biostatistics/1.4.465] [Medline: 12933568]

Abbreviations

HR: health record
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTB: preterm birth
SVM: support vector machine

Edited by C Lovis; submitted 27.09.21; peer-reviewed by M Majurul Ahsan, A Metcalfe; comments to author 04.12.21; revised version received 29.01.22; accepted 26.02.22; published 20.04.22.

Please cite as:
Sharifi-Heris Z, Laitala J, Airola A, Rahmani AM, Bender M
Machine Learning Approach for Preterm Birth Prediction Using Health Records: Systematic Review
JMIR Med Inform 2022;10(4):e33875
URL: https://medinform.jmir.org/2022/4/e33875
doi:10.2196/33875
PMID:35442214

Review

Research and Application of Artificial Intelligence Based on Electronic Health Records of Patients With Cancer: Systematic Review

Xinyu Yang[1,2], PhD; Dongmei Mu[1,2], PhD; Hao Peng[1,2], MSc; Hua Li[1,2], PhD; Ying Wang[1,2], MPH; Ping Wang[1,2], PhD; Yue Wang[1,2], MPH; Siqi Han[1,2], MPH

[1]Division of Clinical Research, The First Hospital of Jilin University, Changchun, China

[2]Department of Medical Informatics, School of Public Health, Jilin University, Changchun, China

Corresponding Author:
Dongmei Mu, PhD
Division of Clinical Research
The First Hospital of Jilin University
No.1, Xinmin Street
Changchun, 130021
China
Phone: 86 0431 81875404
Email: moudm@jlu.edu.cn

Abstract

Background: With the accumulation of electronic health records and the development of artificial intelligence, patients with cancer urgently need new evidence of more personalized clinical and demographic characteristics and more sophisticated treatment and prevention strategies. However, no research has systematically analyzed the application and significance of artificial intelligence based on electronic health records in cancer care.

Objective: The aim of this study was to conduct a review to introduce the current state and limitations of artificial intelligence based on electronic health records of patients with cancer and to summarize the performance of artificial intelligence in mining electronic health records and its impact on cancer care.

Methods: Three databases were systematically searched to retrieve potentially relevant papers published from January 2009 to October 2020. Four principal reviewers assessed the quality of the papers and reviewed them for eligibility based on the inclusion criteria in the extracted data. The summary measures used in this analysis were the number and frequency of occurrence of the themes.

Results: Of the 1034 papers considered, 148 papers met the inclusion criteria. Cancer care, especially cancers of female organs and digestive organs, could benefit from artificial intelligence based on electronic health records through cancer emergencies and prognostic estimates, cancer diagnosis and prediction, tumor stage detection, cancer case detection, and treatment pattern recognition. The models can always achieve an area under the curve of 0.7. Ensemble methods and deep learning are on the rise. In addition, electronic medical records in the existing studies are mainly in English and from private institutional databases.

Conclusions: Artificial intelligence based on electronic health records performed well and could be useful for cancer care. Improving the performance of artificial intelligence can help patients receive more scientific-based and accurate treatments. There is a need for the development of new methods and electronic health record data sharing and for increased passion and support from cancer specialists.

(JMIR Med Inform 2022;10(4):e33799) doi:10.2196/33799

KEYWORDS

electronic health records; artificial intelligence; neoplasms; machine learning

Introduction

Overview

Cancer is known as one of the greatest challenges in health care, and its burden has risen in recent years, calling for a better understanding of clinical prediction strategies in real patient populations. Electronic health records (EHRs) integrate true information about patient care, such as demographics, medical history, and insurance [1]. The secondary use of EHRs is opening immense research avenues and opportunities for improving cancer management. However, there are many challenges of the secondary use of EHRs, and much valuable information is locked behind these vast amounts of complex data. Artificial intelligence (AI) techniques and methods are believed to be the most critical tool to alleviate this issue. Further, an increasing amount of data available in EHRs provides a new environment for the application of AI [2]. With the help of AI-based EHRs, each patient with cancer is more likely to be treated according to the best available knowledge, which is constantly updated for the benefit of the next patient, thereby improving clinical decision-making [3,4]. Despite the rapid development of technology, significant challenges remain to obtain valuable information quickly and accurately based on EHRs to better inform clinical decision-making.

Objectives

The aim of this study was to conduct a review to introduce the current state and limitations of AI based on EHRs from patients with cancer and to explore the opportunities and challenges in this field. The objectives were to review the aspects of categorization of neoplasms, methods and algorithms, and applications in the field of cancer care, EHR data and data sets. These aspects were analyzed to summarize the performance of AI in mining EHRs and its impact on cancer care.

Methods

Search Strategy

The Web of Science Core Collection, PubMed, and the Association for Computing Machinery Digital Library databases were systematically searched to extract potentially relevant papers published from January 2009 to October 2020. The search expression was designed around 3 concepts: AI, cancer, and EHRs. They were combined using the AND Boolean operator. The Web of Science Core Collection search included the following terms, which were selected by referring to the entry terms of Medical Subject Headings and translated for the other databases. The English language was used as an additional filter.

1. AI: AI OR artificial intelligence OR natural language processing OR NLP OR natural language understanding OR NLU OR machine learning OR deep learning OR neural network OR support vector machine OR prediction network OR forecast model OR data mining OR supervised learning OR time series prediction OR intelligence, artificial OR computational intelligence OR intelligence, computational OR machine intelligence OR intelligence, machine OR computer reasoning OR reasoning, computer OR computer vision system OR system, computer vision
2. EHRs: EMR OR electronic medical records OR EHR OR electronic human records OR medical record, electronic OR health record, electronic OR medical record, computerized OR computerized medical record.
3. Cancer: cancer OR oncology OR tumor OR neoplasm OR neoplasia OR tumor OR malignancy

Study Selection

We followed the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines [5]. The abstracts and titles were independently evaluated by 2 reviewers (XY and HP). Two reviewers (XY and PW) independently reviewed the full texts. Reviewers resolved disagreements by reaching consensus and consulted HL after group discussion if they held different opinions. Papers were included in this review if they met the following criteria: (1) peer-reviewed studies only, (2) the studies were on patients with cancer or on solving cancer problems, (3) the research methods used AI, (4) the study data were EHRs and the purpose of the paper was not to build an electronic medical record system, (5) a journal paper or a proceeding paper, (6) a research paper and not a review (including systematic review, meta-analysis, etc), and (7) published in the English language. All reviewers had medical informatics expertise; a basic understanding of EHRs, AI, and cancer; and strict adherence to the inclusion criteria.

Data Collection Process and Data Items

The included papers were cited in an Excel spreadsheet by the reviewers. Reviewers agreed in a group meeting on what to look for in full-texts. According to the research objectives, we retrieved the following data from the key information: study details (including title, author, journal, time of publication), EHR details (including data period, data type, number of sources of data, data set size, data set publicly available, language, patient sample size), AI details (including algorithm categories, precision, negative predictive value, sensitivity [recall], specificity, F-score, accuracy, area under the curve [AUC], and applications), and cancer category. The notes were discussed in a consensus meeting between 2 reviewers after they independently retrieved the detailed data about the items, and they were asked to identify possible bias [6,7] in each paper. Publication bias, unblinded trial bias, and time lag bias were identified. No paper was discarded because of bias. The summary measures used in this analysis were the number and frequency of occurrence of the themes identified by the reviewers. Owing to the heterogeneity in the population, index method [8], and outcomes, we did not perform a quantitative synthesis of the results.

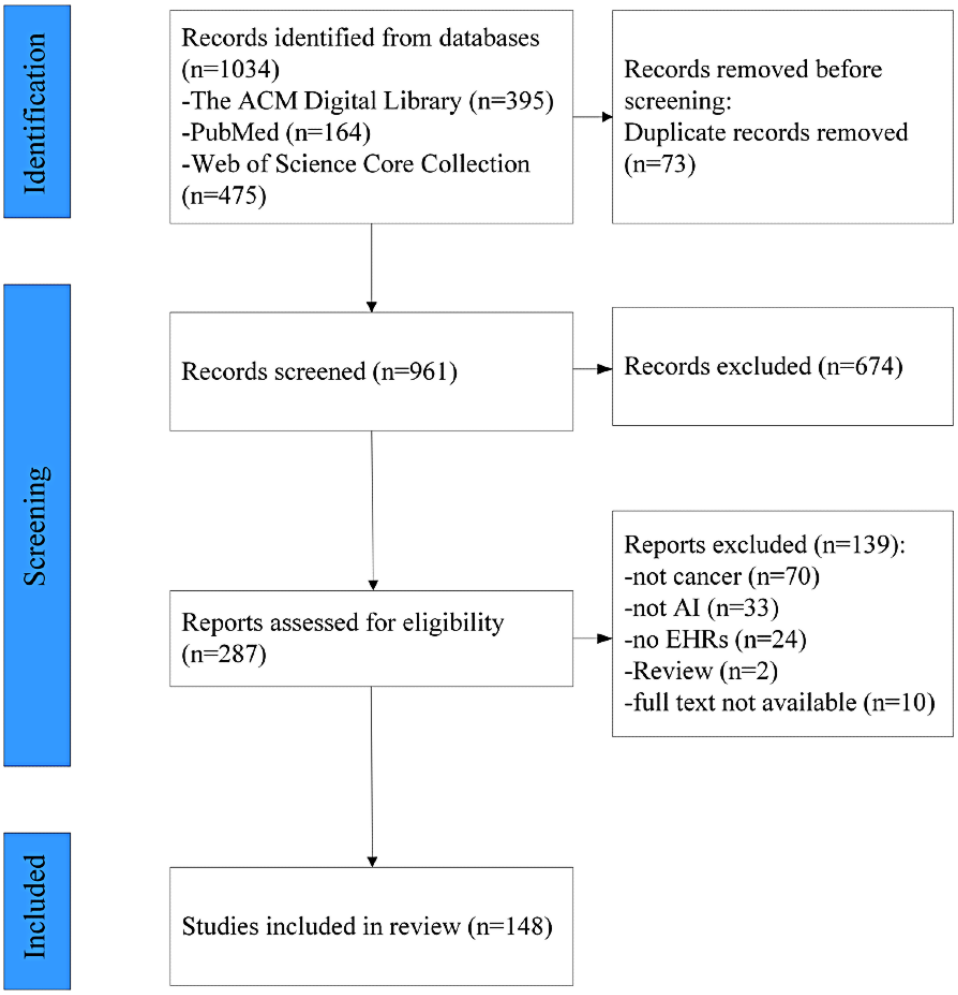
Results

Search and Selection Results

A total of 1034 papers were initially retrieved, with 395 papers from the Association for Computing Machinery Digital Library, 164 from PubMed, and 475 from Web of Science Core Collection; 674 were removed after scanning the titles and abstracts and after removing 73 duplicates; and 287 papers were

ultimately identified for full-text review. Following screening and eligibility, 148 papers were included in the final review. The flowchart of the selection process is presented in Figure 1. The most common reasons for exclusion were as follows: (1) the paper was not directly related to cancer (n=346), (2) the paper was a review and neither a journal paper nor a proceeding paper (n=256), (3) the paper was not based on EHRs (n=134), and (4) the research methods did not incorporate AI (n=67). The observations from each paper are summarized in the spreadsheet shown in Multimedia Appendix 1.

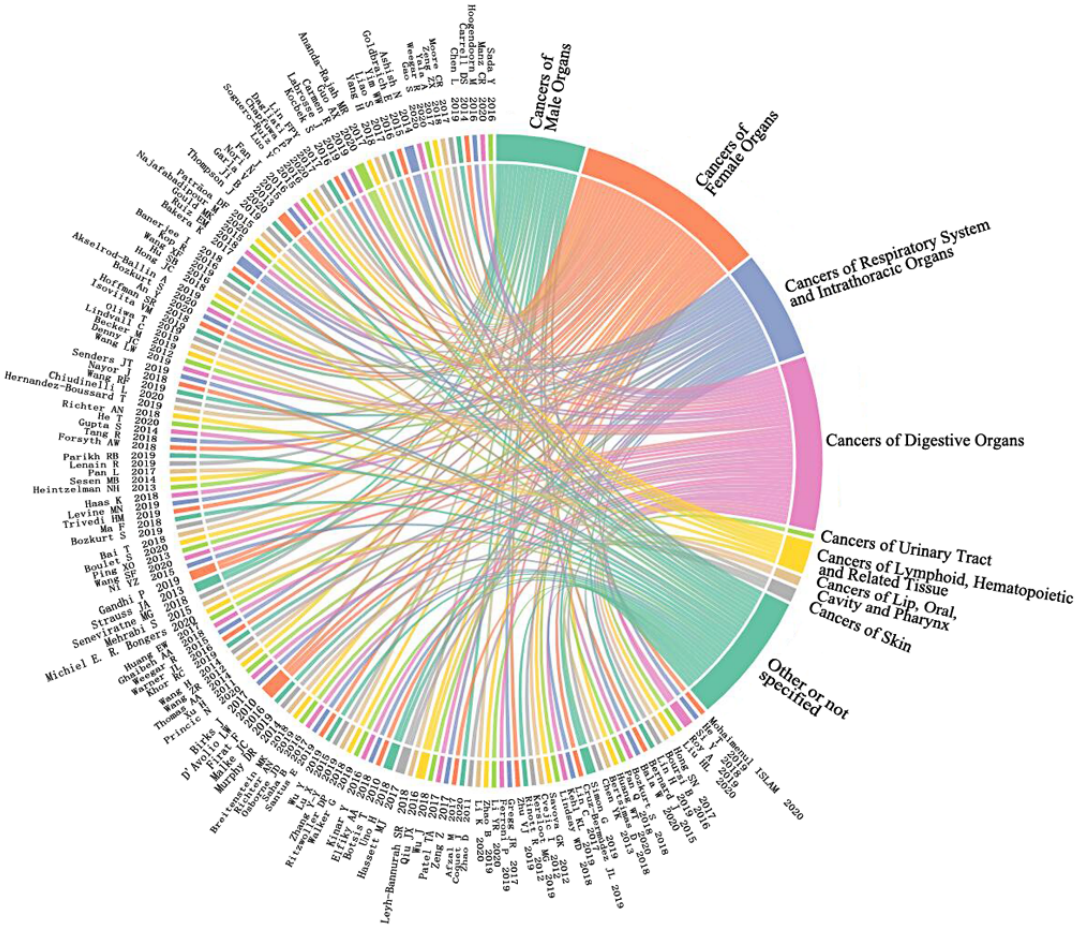
Figure 1. Paper selection flowchart. ACM: Association for Computing Machinery; AI: artificial intelligence; EHR: electronic health record.



Categorization of the Neoplasms

The diseases studied in the 148 papers could be grouped into 9 unique categories of neoplasms according to the anatomical site of the lesion and International Classification of Diseases, tenth revision. The 3 most studied cancer categories were (1) cancers of female organs (n=42), (2) cancers of digestive organs (n=38), and (3) cancers of the respiratory system and intrathoracic organs (n=23). The relationship between each paper and the cancers studied is shown in Figure 2. The complete reference details of the papers cited in Figure 2 are provided in Multimedia Appendix 1. Most of the works on cancers of female organs focused on breast cancer. Receptor status phenotypes, biomarker status, and frequent patterns of care were obtained from EHRs of patients with breast cancer by using AI. For cancers of digestive organs, the types of cancers studied were relatively diverse, mainly comprising colorectal cancer (CRC) and liver cancer types. Earlier detection of CRC attracted the greatest attention from researchers. Because CRC symptoms develop slowly and insidiously over years, early diagnosis offers great opportunity to improve outcomes [9]. AI was constructed to identify the risk of CRC based on demographic and behavioral factors, analysis of complete blood counts [10], and so on. Clinically relevant features of liver cancer were extracted from EHRs, such as tumor reference resolution, tumor number, and largest tumor sizes [11]. Lung cancer was the only cancer of the respiratory system and intrathoracic organs studied in the papers we investigated. For example, a Lung Cancer Assistant was designed to provide decision support for experts in lung cancer multidisciplinary teams [12].

Figure 2. Relationship between the categorizations of the neoplasms and the papers included in this review (the complete reference details of the papers cited in this figure are provided in Multimedia Appendix 1).

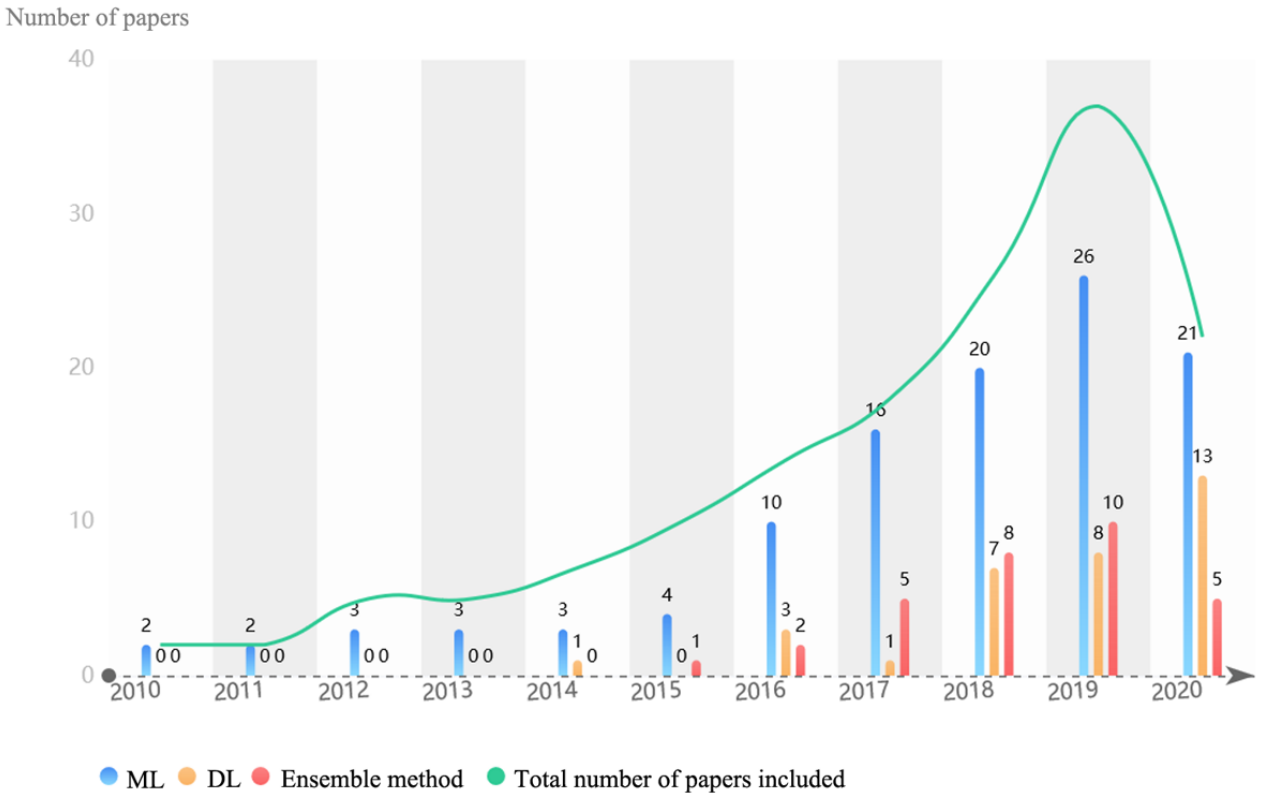


Methods and Algorithms

Machine Learning Algorithms

Machine learning (ML) is an important way to achieve AI. A total of 110 papers used ML algorithms, among which support vector machine (SVM) (n=29) and logistic regression (n=28) were the most commonly used. SVM works well for data sets that are not linearly separable or highly unbalanced, which is important for EHR analysis. Several studies combined SVM with natural language processing (NLP) to extract breast cancer, CRC, and other cancer information from EHRs [13,14]. Logistic regression has been improved to the level of a more sophisticated algorithm for EHR mining of cancer patient data and combined with the lasso penalty [15], a convolutional neural network [16], and other methods in recent years. These algorithms are simple insightful white-box classification algorithms with advantages in interpretability [17] and sensitivity of data details [18]. In fact, these single-model methods were rarely used independently for prediction but used as a baseline to compare the performance of new technologies and methods. However, the deep learning (DL) algorithm and ensemble methods are increasing rapidly (as shown in Figure 3). The ensemble method (n=31), a single strong model combined with multiple weak models, showed high accuracy in processing EHRs. Gradient boosting and random forest performed better than SVM, decision tree, and lasso in classifying free-text pathology reports for prostate cancer into stage groups and identifying cases of metastatic prostate cancer [19,20]. DL (n=33) demonstrated great performance in cancer domains as well. Gao et al [21] designed a modular component with recurrent neural network, including long short-term memory and gated recurrent units for capturing case-level context, to improve the classification accuracy of aggregate-level labels for cancer pathology reports. Recurrent neural network was designed particularly to deal with temporal data, which is very promising for EHRs with timestamps [22]. Qiu et al [23] used convolutional neural network joint training by transferring learning across primary cancer sites to achieve great performance in lung cancer and breast cancer classification tasks. However, these complex and efficient models tend to be black boxes and lack interpretability [24] and transparency, which makes doctors reluctant to accept them. Fortunately, in the papers we reviewed, there have been several attempts to solve this problem, such as the application of attention mechanism [25] and Gradient Class Activation Maps algorithm, decision-making process visualization [26]. In addition, some of the papers in this review have developed novel EHR mining algorithms that perform better than baseline algorithms, such as the “semi-supervised set covering machine” [27] and an unsupervised framework of “subgraph augmented non-negative tensor factorization” [28].

Figure 3. Machine learning algorithms for cancer. DL: deep learning; ML: machine learning.



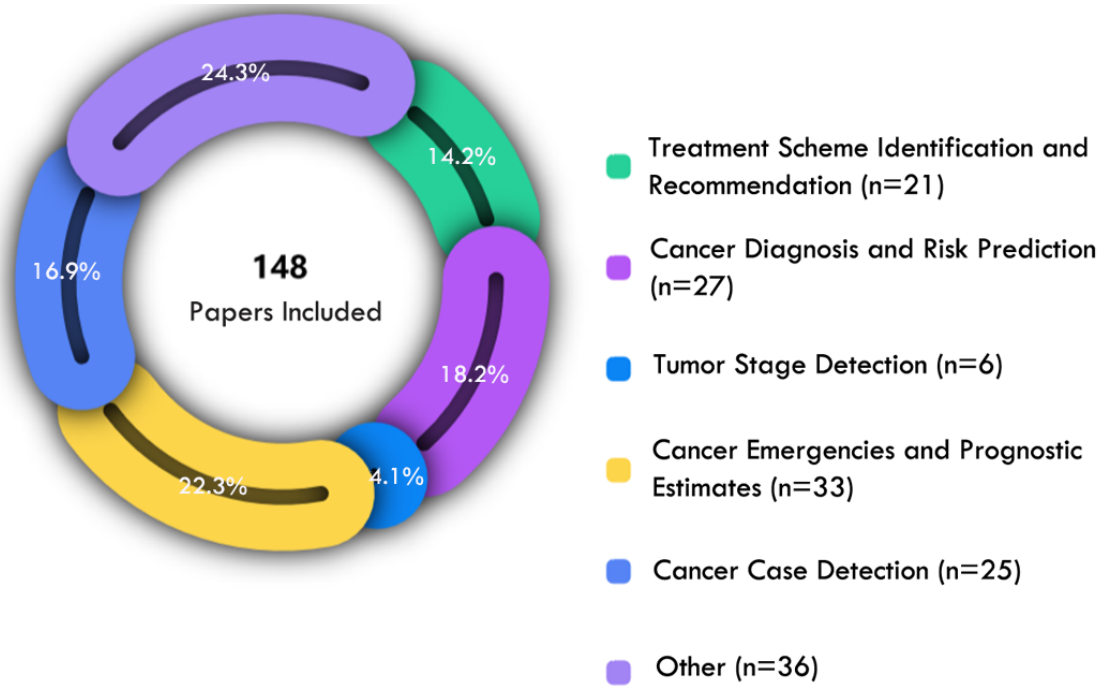
AI Performance Metrics

In our review, 124 papers used one or more of the precision, sensitivity (recall), specificity, F-score, accuracy, and AUC to measure the performance of AI model. The AUC was generally high, that is, 0.7 and above. Accuracy ranged from 0.613 to 1. The precision ranged from 0.353 to 0.999, except for the 4 prediction models for CRC reported by Kop et al [29], Hoogendoorn et al [30], Hong et al [31], and Birks et al [9], wherein their models had precision less than 0.1. Kop et al [29] and Hoogendoorn et al [30] also reported the lowest F-score of 0.058 and 0.074 in this survey, while Ping et al [32] reported the highest F-score of 0.996. Of the papers reporting sensitivity or specificity, 87% had a sensitivity or specificity greater than 0.7 and more than 50% had a sensitivity or specificity greater than 0.9.

Application in the Field of Cancer Care

AI based on EHRs has permeated the whole cycle of cancer care. The significance of the included papers in the journey of cancer medical care can be broadly divided into several applications. The proportion and number of papers showing the application of AI in cancer care are shown in Figure 4. In this section, we summarize the representative papers.

Figure 4. Papers related to artificial intelligence application in cancer care.



Cancer Diagnosis and Risk Prediction

Of the 148 studies, 27 (18.2%) explored the risk factors for cancer, developmental risk prediction models, and differential diagnosis of cancer, maintaining an AUC of 0.7 and above. In the data of 25,430 patients in the United Kingdom, full blood count indicators were added on the basis of age and sex to predict risk of CRC, and it was found that the AUC of the prediction model (based on logistic regression algorithm) at 18-24 months before diagnosis could reach 0.776 [9]. The prostate-specific antigen density, transversal diameter of the prostate, and other variables were used to establish the decision tree model (the variable with maximum gain was selected as the split variable; other hyperparameters used the default settings) to differentiate prostate cancer from benign prostatic hyperplasia [33], achieving a precision of 0.86.

Tumor Stage Detection

Of the 148 studies, 6 (4.1%) used AI to identify explicit and implicit stage information from unstructured EHRs. The performance metrics values of the reported AI models were greater than 0.66. It took less than 1 hour to extract cancer summary stage information from more than 750,000 documents that required a human reader months to years to digest [34]. Two papers explored the staging of lung and prostate cancer with reference to the American Joint Committee on Cancer staging system. Three studies on liver cancer staging used American Joint Committee on Cancer, Barcelona Clinic Liver Cancer, and Cancer of Liver Italian Program staging system.

Treatment Scheme Identification and Recommendation

Of the 148 studies, 21 (14.2%) used AI to adapt doses in antidrug regimen [35], assess effect and combination of dose, evaluate cancer therapeutic procedures, and recommend treatment schemes based on EHRs. The precision, recall, specificity, F-score, accuracy, and AUC were above 0.67, except in a model for drug repurposing reported by Wu et al [36]. Savova et al [37] tried to mine endocrine breast cancer drug treatment patterns by combining information extracted from clinical free text through NLP with structured data, and they obtained high specificity above 0.96 for all categories. Goldbraich et al [38] applied NLP techniques to characterize deviations from clinical practice guidelines in adult soft tissue sarcoma across thousands of patient records, identified that approximately half of all treatment programs deviated from the clinical practice guidelines, and analyzed reasons that may reflect the physicians' rationale in deviation cases. The Oncology Expert Advisor [39] was designed to recommend treatment options by developing a learning model to predict appropriate therapy options for lung cancer with a recall of 0.999, precision of 0.88, and ability to accommodate addition or changes to the approved therapies list.

Cancer Case Detection

Of the 148 studies, 25 (16.9%) proposed AI methods to identify patients with specific cancers such as prostate cancer and breast cancer. The AUC was high above 0.9. Features were extracted from progress notes and pathology reports by NLP, which were used to train the SVM model to identify the group of patients with contralateral breast cancer, obtaining an AUC score of 0.93 (hyperparameters were tuned by 5-fold cross-validation) [40]. The accumulation of EHRs and the development of AI have made it possible to have a large cohort study for different clinical problems. Data-driven intelligent approaches, rather than manual chart review, were important for capturing special cases of cancer among a large cohort efficiently.

Cancer Emergencies and Prognostic Estimates

Of the 148 studies, 33 (22.3%) focused on extracting tumor prognostic factors, predicting outcomes in individual patients with cancer and developing emergency prediction models for emergency visits and hospital admissions and so on. All reported

AUCs were greater than 0.72. Gradient tree boosting model [41] was developed to predict emergency visits and hospital admissions during radiation and chemoradiation based on synthesizing and processing EHRs (demographics, drug therapy, etc) with an AUC of 0.798 (hyperparameters were tuned by 5-fold cross-validation). Regarding the prediction of cancer relapse, patients [42] with childhood acute lymphoblastic leukemia were classified into different relapse risk-level groups by random forest algorithms based on EHRs (white blood cell count, hemoglobin, etc) with an AUC of more than 0.9. For the prediction of cancer survival, breast cancer–related variables, tumor characteristics, and patient demographics were used to developed SVM models (the soft margin parameter C of SVM was selected through cross-validation) to estimate the patient's survival status of the 3 time periods. AI models were slightly better than the performance of the clinician panel [43]. Compared with traditional methods for survival analysis, AI methods focused on the prediction of event occurrence, applied to high-dimensional problems usually, and showed improvements in predictive performance [44].

Data and Data Sets

Most papers described experiments conducted on non–publicly available data sets, and more than half of the papers were based on data from a single health care institution, as detailed in Multimedia Appendix 1. Less than 10% of the included papers (n=12) made use of publicly available data sets, that is, SEER, Informatics for Integrating Biology and the Bedside, and Medical Information Mart for Intensive Care data set. A few studies combined clinical practice guidelines, a literature corpus, administrative data, and other types of data on the basis of using EHRs. Focusing on the patient sample size used in the actual study and eliminating the remaining 35 papers that were not specified, 42 had fewer than 500 samples, 17 had between 500 and 1000 samples, and only 18 had over 10,000 samples. Regarding the language used in EHRs, 100 papers exploiting EHRs in English topped the list, followed by papers with EHRs in Chinese (n=18). Algorithms for English report processing have been relatively effective and can be scaled to other languages. For example, an NLP algorithm automatically extracting carcinoma and atypia entities from English pathology reports achieved an accuracy of 0.9 [45]. It was later applied to Chinese breast pathology reports. In comparison with using English reports, this paper [46] discussed the performance of the model and demonstrated that it worked just as well for Chinese processing. Regarding the nature and challenges of EHRs used in the experiment, nearly half of the studies explicitly used only unstructured data such as pathology reports, progress notes, discharge notes, and radiology reports.

Discussion

Principal Findings

Of 1034 studies, 148 were selected for the systematic review. Our systematic review has shown that the use of AI to process EHRs has broad applications in providing insights into cancer care, particularly for cancers of female organs, digestive organs, respiratory organs, and intrathoracic organs. ML was the common implementation of AI based on the EHRs of patients with cancer. SVM and logistic regression were the most used ML classifiers. Traditional ML algorithms moved from stand-alone predictions to benchmarks for new approaches. Ensemble methods and DL are on the rise and improving performance. However, the interpretability of complex algorithms is a key issue, and more research is needed on this issue. The results show that most AI models can usually achieve a performance metric value of 0.7. It is worth noting that the CRC prediction models reported in 4 papers had significantly lower precision and 2 of them had lower F-scores. Further investigation revealed that in the design of the experiment, the researchers consciously traded higher false-positive rates for fewer patients that were missed because they believed that the cost of a normal person being wrongly predicted was lower than the cost of missing a patient depending on the characteristics of CRC. However, high false-positive rates would also make medical procedures too costly or invasive and should be analyzed according to the disease investigated. Cancer care could benefit from AI based on EHRs through cancer emergencies and prognostic estimates, cancer diagnosis and prediction, tumor stage detection, cancer case detection, and treatment pattern recognition. The topic of emergency and prognostic estimation had the most research. Finally, we discussed EHRs and databases. Our review found that the vast majority of studies in this area were based on private databases within the institution, resulting in poor portability of the proposed methodology process. Public databases were underused, and few patient records were included in the actual studies. In another way, it also reflects the fact that public databases are still scarce. English EHRs are mainly used, and the exploration of EHRs in other languages is limited. Of course, this may be a bias caused by our selection of English papers only. Fortunately, the existing literature also showed that the processing methods of EHRs in English are relatively mature, and these methods may be transplanted to data in other languages. Much cancer information are stored in unstructured formats of EHRs and are difficult to mine, thus requiring better algorithms and more efforts. Furthermore, EHRs can be combined with other data sources to support AI for cancer care.

Comparison With Prior Work

Recently, several systematic reviews related to EHRs have been published, with particular attention given to the implementation of EHR systems [47,48]. Several studies have discussed different applications of technology to EHRs, such as blockchain [49]; yet, few have focused on the specific secondary use of EHRs, such as the role in reducing unwarranted clinical variation [6] and patient identification and clinical support in palliative care [50], with even fewer focusing on specific disease areas such as diabetes [51]. There is existing work elucidating the state of AI research in cancers [52,53]. However, to our knowledge, none have focused specifically on the combination of EHRs and AI in cancer, which makes it difficult to have a specific understanding of the current implementation and challenges of this field.

Limitations

This review examined nearly 12 years of literature and may have the following limitations. First, despite efforts to develop

a systematic and careful search strategy, there is no guarantee that all relevant literature will be included. Our search was limited to published literature in English, but searches in other languages or gray literature may provide additional findings. Second, the popularity of EHRs and the degree of data development vary in different countries and environments, which may lead to inconsistency in the quality of the included literature research, and the algorithms and effect evaluation analysis may have an impact. Third, we only considered the literature and did not investigate the AI products in the market. This may need to be further supplemented.

Conclusions

Our review shows that AI based on EHRs performed well and can be useful for cancer care in 4 areas: categorization of neoplasms, methods and algorithms, application in the field of cancer care, and data and data sets. Based on our review, we propose the following recommendations for future research:

1. The development of new AI methods: The use of hybrid approaches could improve the performance of AI models. DL and ensemble methods have great potential in cancer care. The interpretability of methods must be given more attention. In addition, the models need to adjust the evaluation of performance appropriately according to the disease under study so that it can achieve better practical results.
2. EHR sharing and fusion: There are too few open data sets available for researchers, and the lack of a large annotated gold standard library has become a major bottleneck for research in this field. In the case of complying with data ethics, the sharing of EHRs and multiagency participation in EHR databases is urgently needed. Guidelines, literature data, and corpora in other fields can play an important role in addressing this problem. At the same time, EHRs could be complemented by guides, literature, and corpora in other fields to enhance the benefits of AI.
3. Passion and support from cancer specialists: Recognition and acceptance by practitioners in the fields of cancer care is necessary for the research results to be translated to practice. This requires more human experts in this field to overcome the natural resistance of traditional views, participate in the formulation of a gold standard, reasonably adopt research conclusions, and take responsibility for the actual medical outcomes.

Acknowledgments

This work was supported by grant awards from the National Natural Science Foundation of China (grant 71974074), the Jilin Scientific and Technological Development Program (grant 20200301004RQ), and the 2021 Higher Education Scientific Research Project of Jilin Association for Higher Education (grant JGJX2021C3).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete list of the reviewed papers.
[XLSX File (Microsoft Excel File), 53 KB - medinform_v10i4e33799_app1.xlsx]

References

1. Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. Cell 2019 Mar 21;177(1):58-69 [FREE Full text] [doi: 10.1016/j.cell.2019.02.039] [Medline: 30901549]
2. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The Evolving Use of Electronic Health Records (EHR) for Research. Semin Radiat Oncol 2019 Oct;29(4):354-361. [doi: 10.1016/j.semradonc.2019.05.010] [Medline: 31472738]
3. Tenenbaum JM, Shrager J. Cancer: A Computational Disease that AI Can Cure. AIMag 2011 Jun 05;32(2):14. [doi: 10.1609/aimag.v32i2.2345]
4. Fessele KL. The Rise of Big Data in Oncology. Semin Oncol Nurs 2018 May;34(2):168-176. [doi: 10.1016/j.soncn.2018.03.008] [Medline: 29606536]
5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. J Clin Epidemiol 2021 Jun;134:178-189 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.03.001] [Medline: 33789819]
6. Hodgson T, Burton-Jones A, Donovan R, Sullivan C. The Role of Electronic Medical Records in Reducing Unwarranted Clinical Variation in Acute Health Care: Systematic Review. JMIR Med Inform 2021 Nov 17;9(11):e30432 [FREE Full text] [doi: 10.2196/30432] [Medline: 34787585]
7. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions version 6.2 (updated February 2021). Cochrane. URL: https://training.cochrane.org/handbook [accessed 2022-01-20]

8. Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, et al. Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review. JMIR Med Inform 2021 Sep 09;9(9):e30401 [FREE Full text] [doi: 10.2196/30401] [Medline: 34499041]
9. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med 2017 Oct;6(10):2453-2460 [FREE Full text] [doi: 10.1002/cam4.1183] [Medline: 28941187]
10. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. J Am Med Inform Assoc 2016 Sep;23(5):879-890 [FREE Full text] [doi: 10.1093/jamia/ocv195] [Medline: 26911814]
11. Yim W, Kwan SW, Yetisgen M. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. J Biomed Inform 2016 Dec;64:179-191 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.005] [Medline: 27729234]
12. Sesen MB, Peake MD, Banares-Alcantara R, Tse D, Kadir T, Stanley R, et al. Lung Cancer Assistant: a hybrid clinical decision support application for lung cancer care. J R Soc Interface 2014 Sep 06;11(98):20140534 [FREE Full text] [doi: 10.1098/rsif.2014.0534] [Medline: 24990290]
13. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc 2011;2011:1564-1572 [FREE Full text] [Medline: 22195222]
14. Kocbek S, Cavedon L, Martinez D, Bain C, Manus CM, Haffari G, et al. Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. J Biomed Inform 2016 Dec;64:158-167 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.008] [Medline: 27742349]
15. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, et al. Structure-Leveraged Methods in Breast Cancer Risk Prediction. J Mach Learn Res 2016 Dec;17:85 [FREE Full text] [Medline: 28559747]
16. Lin C, Hsu C, Lou Y, Yeh S, Lee C, Su S, et al. Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes. J Med Internet Res 2017 Nov 06;19(11):e380 [FREE Full text] [doi: 10.2196/jmir.8344] [Medline: 29109070]
17. Bertsimas D, Dunn J, Pawlowski C, Silberholz J, Weinstein A, Zhuo YD, et al. Applied Informatics Decision Support Tool for Mortality Predictions in Patients With Cancer. JCO Clinical Cancer Informatics 2018 Dec(2):1-11. [doi: 10.1200/cci.18.00003]
18. Lindsay WD, Ahern CA, Tobias JS, Berlind CG, Chinniah C, Gabriel PE, et al. Automated data extraction and ensemble methods for predictive modeling of breast cancer outcomes after radiation therapy. Med Phys 2019 Feb;46(2):1054-1063. [doi: 10.1002/mp.13314] [Medline: 30499597]
19. Lenain R, Seneviratne MG, Bozkurt S, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine Learning Approaches for Extracting Stage from Pathology Reports in Prostate Cancer. Stud Health Technol Inform 2019 Aug 21;264:1522-1523 [FREE Full text] [doi: 10.3233/SHTI190515] [Medline: 31438212]
20. Seneviratne MG, Banda JM, Brooks JD, Shah NH, Hernandez-Boussard TM. Identifying Cases of Metastatic Prostate Cancer Using Machine Learning on Electronic Health Records. AMIA Annu Symp Proc 2018;2018:1498-1504 [FREE Full text] [Medline: 30815195]
21. Gao S, Alawad M, Schaefferkoetter N, Penberthy L, Wu X, Durbin EB, et al. Using case-level context to classify cancer pathology reports. PLoS One 2020;15(5):e0232840 [FREE Full text] [doi: 10.1371/journal.pone.0232840] [Medline: 32396579]
22. Yadav P, Steinbach M, Kumar V, Simon G. Mining Electronic Health Records (EHRs). ACM Comput. Surv 2018 Nov 30;50(6):1-40. [doi: 10.1145/3127881]
23. Qiu JX, Yoon H, Fearn PA, Tourassi GD. Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports. IEEE J Biomed Health Inform 2018 Jan;22(1):244-251. [doi: 10.1109/JBHI.2017.2700722] [Medline: 28475069]
24. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760 [FREE Full text] [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]
25. Bai T, Egleston BL, Zhang S, Vucetic S. Interpretable Representation Learning for Healthcare via Capturing Disease Progression through Time. KDD 2018 Aug;2018:43-51 [FREE Full text] [doi: 10.1145/3219819.3219904] [Medline: 31037221]
26. Bala W, Steinkamp J, Feeney T, Gupta A, Sharma A, Kantrowitz J, et al. A Web Application for Adrenal Incidentaloma Identification, Tracking, and Management Using Machine Learning. Appl Clin Inform 2020 Aug;11(4):606-616 [FREE Full text] [doi: 10.1055/s-0040-1715892] [Medline: 32937677]
27. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One 2012;7(1):e30412 [FREE Full text] [doi: 10.1371/journal.pone.0030412] [Medline: 22276193]

28. Luo Y, Xin Y, Hochberg E, Joshi R, Uzuner O, Szolovits P. Subgraph augmented non-negative tensor factorization (SANTF) for modeling clinical narrative text. J Am Med Inform Assoc 2015 Sep;22(5):1009-1019 [FREE Full text] [doi: 10.1093/jamia/ocv016] [Medline: 25862765]
29. Kop R, Hoogendoorn M, Teije AT, Büchner FL, Slottje P, Moons LM, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med 2016 Sep 01;76:30-38. [doi: 10.1016/j.compbiomed.2016.06.019] [Medline: 27392227]
30. Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med 2016 May;69:53-61 [FREE Full text] [doi: 10.1016/j.artmed.2016.03.003] [Medline: 27085847]
31. Hong SN, Son HJ, Choi SK, Chang DK, Kim Y, Jung S, et al. A prediction model for advanced colorectal neoplasia in an asymptomatic screening population. PLoS One 2017;12(8):e0181040 [FREE Full text] [doi: 10.1371/journal.pone.0181040] [Medline: 28841657]
32. Ping X, Tseng Y, Chung Y, Wu Y, Hsu C, Yang P, et al. Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. Telemed J E Health 2013 Sep;19(9):704-710. [doi: 10.1089/tmj.2012.0241] [Medline: 23869395]
33. Zhang Y, Li Q, Xin Y, Lv W. Differentiating Prostate Cancer from Benign Prostatic Hyperplasia Using PSAD Based on Machine Learning: Single-Center Retrospective Study in China. IEEE/ACM Trans. Comput. Biol. and Bioinf 2019 May 1;16(3):936-941. [doi: 10.1109/tcbb.2018.2822675]
34. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. ReCAP: Feasibility and Accuracy of Extracting Cancer Stage Information From Narrative Electronic Health Record Data. J Oncol Pract 2016 Feb;12(2):157-8; e169. [doi: 10.1200/JOP.2015.004622] [Medline: 26306621]
35. Boulet S, Ursino M, Thall P, Landi B, Lepère C, Pernot S, et al. Integration of elicited expert information via a power prior in Bayesian variable selection: Application to colon cancer data. Stat Methods Med Res 2020 Feb;29(2):541-567 [FREE Full text] [doi: 10.1177/0962280219841082] [Medline: 30963815]
36. Wu Y, Warner JL, Wang L, Jiang M, Xu J, Chen Q, et al. Discovery of Noncancer Drug Effects on Survival in Electronic Health Records of Patients With Cancer: A New Paradigm for Drug Repurposing. JCO Clin Cancer Inform 2019 May;3:1-9 [FREE Full text] [doi: 10.1200/CCI.19.00001] [Medline: 31141421]
37. Savova GK, Olson JE, Murphy SP, Cafourek VL, Couch FJ, Goetz MP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. J Am Med Inform Assoc 2012 Jun;19(e1):e83-e89 [FREE Full text] [doi: 10.1136/amiajnl-2011-000295] [Medline: 22140207]
38. Goldbraich E, Waks Z, Farkash A, Monti M, Torresani M, Bertulli R, et al. Understanding Deviations from Clinical Practice Guidelines in Adult Soft Tissue Sarcoma. Stud Health Technol Inform 2015;216:280-284. [Medline: 26262055]
39. Simon G, DiNardo CD, Takahashi K, Cascone T, Powers C, Stevens R, et al. Applying Artificial Intelligence to Address the Knowledge Gaps in Cancer Care. Oncologist 2019 Jun;24(6):772-782 [FREE Full text] [doi: 10.1634/theoncologist.2018-0257] [Medline: 30446581]
40. Zeng Z, Li X, Espino S, Roy A, Kitsch K, Clare S, et al. Contralateral Breast Cancer Event Detection Using Nature Language Processing. AMIA Annu Symp Proc 2017;2017:1885-1892 [FREE Full text] [Medline: 29854260]
41. Hong JC, Niedzwiecki D, Palta M, Tenenbaum JD. Predicting Emergency Visits and Hospital Admissions During Radiation and Chemoradiation: An Internally Validated Pretreatment Machine Learning Algorithm. JCO Clinical Cancer Informatics 2018 Dec(2):1-11. [doi: 10.1200/cci.18.00037]
42. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, et al. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep 2017 Aug 07;7(1):7402 [FREE Full text] [doi: 10.1038/s41598-017-07408-0] [Medline: 28784991]
43. Gupta S, Tran T, Luo W, Phung D, Kennedy RL, Broad A, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open 2014 Mar 17;4(3):e004007 [FREE Full text] [doi: 10.1136/bmjopen-2013-004007] [Medline: 24643167]
44. Chapfuwa P, Li C, Mehta N, Carin L, Henao R. Survival cluster analysis. 2020 Presented at: Proceedings of the ACM Conference on Health, Inference, and Learning; 2020; Toronto, Ontario, Canada. [doi: 10.1145/3368555.3384465]
45. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, et al. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 2017 Jan;161(2):203-211. [doi: 10.1007/s10549-016-4035-1] [Medline: 27826755]
46. Tang R, Ouyang L, Li C, He Y, Griffin M, Taghian A, et al. Machine learning to parse breast pathology reports in Chinese. Breast Cancer Res Treat 2018 Jun;169(2):243-250. [doi: 10.1007/s10549-018-4668-3] [Medline: 29380208]
47. Fernández-Alemán JL, Señor IC, Lozoya, Toval A. Security and privacy in electronic health records: a systematic literature review. J Biomed Inform 2013 Jun;46(3):541-562 [FREE Full text] [doi: 10.1016/j.jbi.2012.12.003] [Medline: 23305810]
48. Bernal JL, DelBusto S, García-Mañoso MI, de Castro Monteiro E, Moreno, Varela-Rodríguez C, et al. Impact of the implementation of electronic health records on the quality of discharge summaries and on the coding of hospitalization episodes. Int J Qual Health Care 2018 Oct 01;30(8):630-636. [doi: 10.1093/intqhc/mzy075] [Medline: 29668920]
49. Mayer AH, da Costa CA, Righi RDR. Electronic health records in a Blockchain: A systematic review. Health Informatics J 2020 Jun;26(2):1273-1288 [FREE Full text] [doi: 10.1177/1460458219866350] [Medline: 31566472]

50. Bush RA, Pérez A, Baum T, Etland C, Connelly CD. A systematic review of the use of the electronic health record for patient identification, communication, and clinical support in palliative care. JAMIA Open 2018 Oct 01;1(2):294-303 [FREE Full text] [doi: 10.1093/jamiaopen/ooy028] [Medline: 30842998]
51. Lessing SE, Hayman LL. Diabetes Care and Management Using Electronic Medical Records: A Systematic Review. J Diabetes Sci Technol 2019 Jul;13(4):774-782 [FREE Full text] [doi: 10.1177/1932296818815507] [Medline: 30556418]
52. Akazawa M, Hashimoto K. Artificial intelligence in gynecologic cancers: Current status and future challenges - A systematic review. Artif Intell Med 2021 Oct;120:102164. [doi: 10.1016/j.artmed.2021.102164] [Medline: 34629152]
53. Jin P, Ji X, Kang W, Li Y, Liu H, Ma F, et al. Artificial intelligence in gastric cancer: a systematic review. J Cancer Res Clin Oncol 2020 Sep;146(9):2339-2350. [doi: 10.1007/s00432-020-03304-9] [Medline: 32613386]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
CRC: colorectal cancer
DL: deep learning
EHR: electronic health record
ML: machine learning
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses
SVM: support vector machine

Edited by C Lovis; submitted 24.09.21; peer-reviewed by X Dong, F Taghizadeh-Hesary, Y Chu; comments to author 04.12.21; revised version received 24.01.22; accepted 14.03.22; published 20.04.22.

Please cite as:
Yang X, Mu D, Peng H, Li H, Wang Y, Wang P, Wang Y, Han S
Research and Application of Artificial Intelligence Based on Electronic Health Records of Patients With Cancer: Systematic Review
JMIR Med Inform 2022;10(4):e33799
URL: https://medinform.jmir.org/2022/4/e33799
doi:10.2196/33799
PMID:35442195

Original Paper

Global Scientific Research Landscape on Medical Informatics From 2011 to 2020: Bibliometric Analysis

Xuefei He[1], MD; Cheng Peng[2], MD, PhD; Yingxin Xu[3], BSc; Ye Zhang[3], BSc; Zhongqing Wang[3], PhD

[1]Department of Ophthalmology, Hwa Mei Hospital, University of Chinese Academy of Sciences, Ningbo, China

[2]Department of Ophthalmology, The Fourth Affiliated Hospital of China Medical University, Shenyang, China

[3]Information Center, The First Hospital of China Medical University, Shenyang, China

Corresponding Author:
Zhongqing Wang, PhD
Information Center
The First Hospital of China Medical University
155 Nanjingbei Street
Shenyang, 110001
China
Phone: 86 15940082159
Email: wangzhongqing@cmu.edu.cn

Abstract

Background: With the emerging information and communication technology, the field of medical informatics has dramatically evolved in health care and medicine. Thus, it is crucial to explore the global scientific research landscape on medical informatics.

Objective: This study aims to present a visual form to clarify the overall scientific research trends of medical informatics in the past decade.

Methods: A bibliometric analysis of data retrieved and extracted from the Web of Science Core Collection (WoSCC) database was performed to analyze global scientific research trends on medical informatics, including publication year, journals, authors, institutions, countries/regions, references, and keywords, from January 1, 2011, to December 31, 2020.

Results: The data set recorded 34,742 articles related to medical informatics from WoSCC between 2011 and 2020. The annual global publications increased by 193.86% from 1987 in 2011 to 5839 in 2020. Journal of Medical Internet Research (3600 publications and 63,932 citations) was the most productive and most highly cited journal in the field of medical informatics. David W Bates (99 publications), Harvard University (1161 publications), and the United States (12,927 publications) were the most productive author, institution, and country, respectively. The co-occurrence cluster analysis of high-frequency author keywords formed 4 clusters: (1) artificial intelligence in health care and medicine; (2) mobile health; (3) implementation and evaluation of electronic health records; (4) medical informatics technology application in public health. COVID-19, which ranked third in 2020, was the emerging theme of medical informatics.

Conclusions: We summarize the recent advances in medical informatics in the past decade and shed light on their publication trends, influential journals, global collaboration patterns, basic knowledge, research hotspots, and theme evolution through bibliometric analysis and visualization maps. These findings will accurately and quickly grasp the research trends and provide valuable guidance for future medical informatics research.

(JMIR Med Inform 2022;10(4):e33842) doi:10.2196/33842

KEYWORDS

medical informatics; bibliometrics; VOSviewer; data visualization

Introduction

Background

The field of medical informatics is dedicated to systematically processing data, information, and knowledge in medicine and health care [1]. In the 1950s, Robert S Ledley and Lee Browning Lusted first performed the complicated reasoning processes inherent in medical diagnosis using electronic computers to minimize medical errors primarily [2]. Given that the emerging information and communication technology is being continuously applied in the medical field, medical informatics, as a discipline, has dramatically evolved over the past 70 years

and brought about significant changes to human social needs [2,3]. Recent advances in health care information technology, electronic health records (EHRs), health data standards, and health information exchange have become the major focus of scientific research [4]. Therefore, health informatics, which represents the development of systems and methods for acquisition, processing, handling, communication, storage, retrieval, management, discovery, analyzing, and synthesizing patient information to improve health and health care, is more often used in the literature [5-8]. The number of publications and journals focusing on medical informatics/health informatics has multiplied in recent years. Therefore, it is essential to explore the global scientific research landscape in this discipline.

Bibliometrics is defined as scientific and quantitative research of publications, which describes the research trends of a certain research field using statistical methods to analyze a large number of publications [9]. In 2007, Bansard et al [10] first presented a bibliometric study on medical informatics and bioinformatics, which mainly identified the present links and potential synergies between the bioinformatics and medical informatics research areas. Subsequently, bibliometric analyses on specific medical informatics technology have been performed, such as those on mobile health research [11], shared decision making [12], telemedicine [13-15], computer-aided diagnosis [16], natural language processing [17], artificial intelligence (AI) in health care [18], digital health [19,20], among others.

Objective

This study aims to analyze medical informatics as a discipline (a catalog from the Web of Science Core Collection [WoSCC]) and demonstrate the longitudinal trends from the global perspective. Thus, we performed bibliometric analysis and prepared visualization maps to identify and present the publication trends, global collaboration patterns, basic knowledge, research hotspots, and emerging hotspots in medical informatics.

Methods

Data Collection

WoSCC is the most widely used database in various scientific fields, including over 9000 high-level academic journals worldwide. The research area is generated by a set of classification methods for all databases under WoSCC. Therefore, documents of the same research area or discipline can be identified, retrieved, and analyzed from WoSCC for bibliometrics analysis [21]. Medical informatics is one of the 252 research areas of WoSCC. For search purposes, the retrieval research area was set as “medical informatics”, the period was set as “from 2011 to 2020”, the document type was set as “article”, and the language was set as “English”. We conducted our search strategy in WoSCC on June 1, 2021, at China Medical University.

We identified and incorporated 34,742 studies on medical informatics from WoSCC. The full record and cited references of the retrieved publications were collected and saved in text formats (eg, .txt). The data used in this study are publicly available and associated with no protected health information.

Ethics Consideration

Publicly available data were searched and downloaded from WoSSC. The extraction of this data did not involve interaction with human subjects or animals. Thus, there were no ethical issues involving the use of these data, and no approval from an Ethics Committee was required.

Analytical Tool and Visualization Maps

The most commonly used bibliometric methods are co-authorship, co-citation, and co-occurrence. Co-authorship analysis reveals collaboration patterns among authors, institutions, and countries [22]. Co-citation analysis contributes to discovering and determining the knowledge base of one discipline [23]. Co-occurrence analysis uses the frequency of multiple words in the same article to identify how close they are, thereby helping researchers identify hot topics and trends in the discipline. VOSviewer [24] is an excellent bibliometric analysis software developed by van Eck and Waltman [25,26]. It calculates the similarity s_{ij} of 2 items i and j with the equation $s_{ij} = c_{ij}/(w_i w_j)$, where c_{ij} denotes the number of co-occurrences of items i and j, and w_i and w_j denote the total number of occurrences of items i and j. Once the similarity matrix is created, VOSviewer maps all the items in a 2D map so that items with a high similarity will be located close to each other, while those with a low similarity will be located far from each other. In this study, we employed VOSviewer version 1.6.16 to extract bibliometric information such as publication year, journals, authors, institutions, countries/regions, references, and keywords. Besides, we employed VOSviewer to conduct co-authorship analysis, co-citation analysis, co-occurrence analysis, and then built visualization network maps.

Results

Global Publications on Medical Informatics

A total of 34,742 articles on medical informatics were retrieved. The average annual number of publications was 3474 during the past decade. The annual global publications on entire life sciences and biomedical sciences are presented in Multimedia Appendix 1. The annual global publications on entire life sciences and biomedical sciences increased 55.73% from 573,981 to 893,887 from 2011 to 2020. The annual global publications on medical informatics increased 193.86% from 1987 to 5839 from 2011 to 2020, which was the highest increase rate in the life sciences and biomedical fields.

The global distribution of countries/regions participating in medical informatics research is shown in Figure 1. A total of 161 countries/regions contributed to medical informatics from 2011 to 2020. The top 10 countries contributed 27,213 articles in medical informatics. The United States (12,927 publications) is the most productive country, followed by Germany (3336 publications), England (3269 publications), China (3157 publications), and Canada (2237 publications). Changes in the annual ranking of the top 10 most productive countries for medical informatics research are shown in Figure 2. The rankings of the top 10 countries changed every year from 2011 to 2020, but the United States consistently ranked first in publications.

Figure 1. The global distribution of countries/regions participating in medical informatics research from 2011 to 2020.

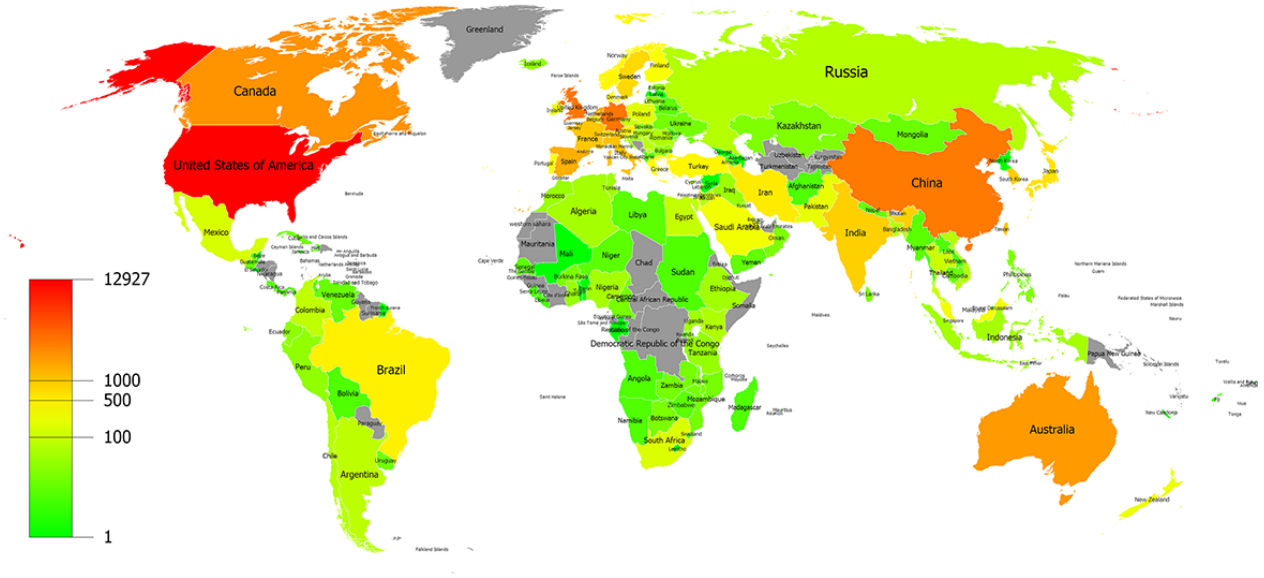
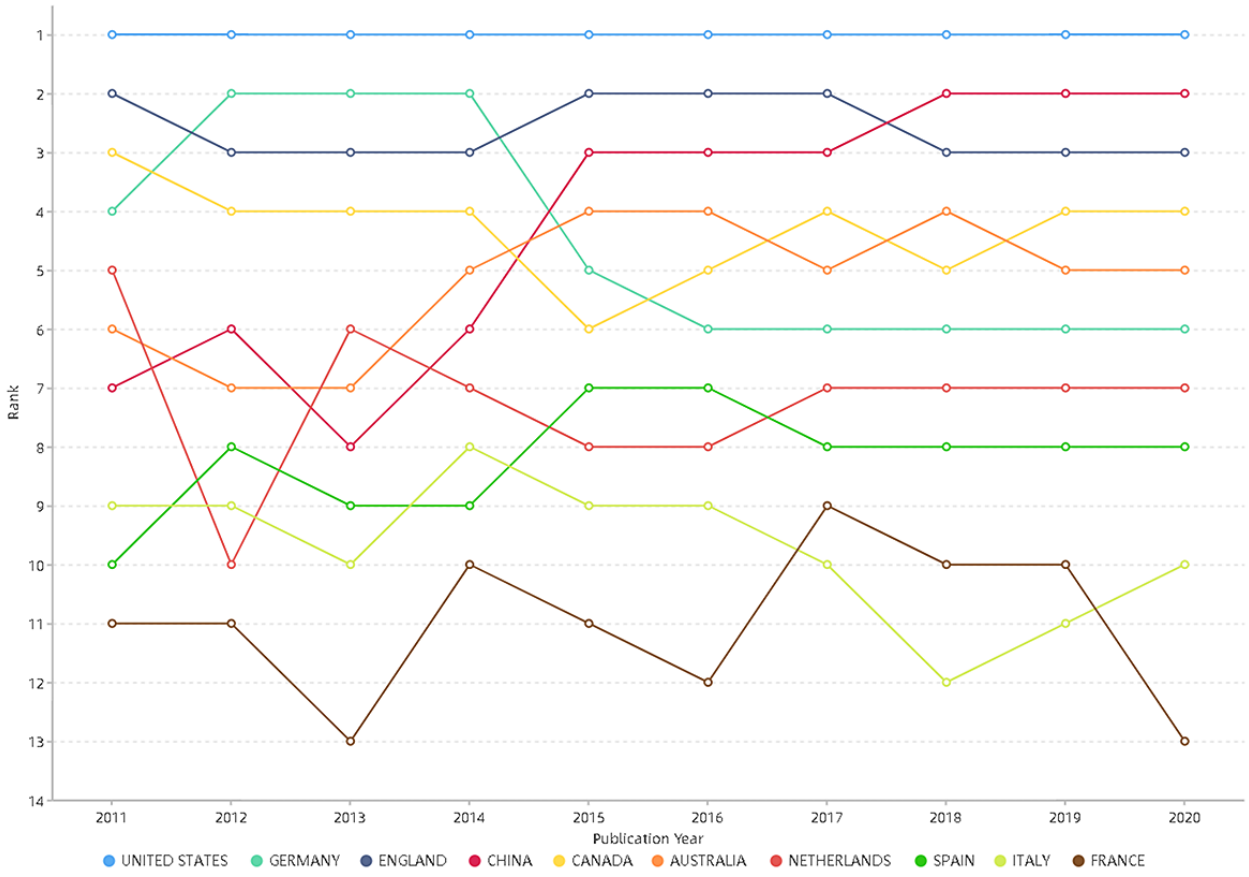


Figure 2. The annual ranking changes of the top 10 most productive countries regarding publication of articles on medical informatics from 2011 to 2020.



Contribution of Source Journals

Based on the retrieved results, articles on medical informatics were distributed in 37 journals. The top 10 journals with the most publications in the medical informatics discipline are presented in Table 1. From 2011 to 2020, *Journal of Medical Internet Research* with 3600/34,742 (10.36%) publications was the top productive journal, followed by *Statistics in Medicine* (3282 publications) and *Computer Methods and Programs in Biomedicine* (2409 publications). *Journal of Medical Internet Research* with 63,932 citations was also the most highly cited journal, followed by *Statistics in Medicine* (45,042 citations) and *Journal of the American Medical Informatics Association* (36,874 citations).

Table 1. The top 10 most productive journals in Medical Informatics from 2011 to 2020.

Rank	Journal	Country	Publication start year	Total publications	Total citations
1	*Journal of Medical Internet Research*	Canada	1999	3600	63,932
2	*Statistics in Medicine*	England	1982	3282	45,042
3	*Computer Methods and Programs in Biomedicine*	Ireland	1985	2409	35,157
4	*Biomedical Engineering-Biomedizinische Technik*	Germany	1971	2184	3541
5	*Journal of Medical Systems*	United States	1977	2104	28,747
6	*Journal of The American Medical Informatics Association*	England	1994	1746	36,874
7	*Journal of Evaluation in Clinical Practice*	England	1995	1676	13,720
8	*BMC Medical Informatics and Decision Making*	England	2001	1660	17,956
9	*IEEE Journal of Biomedical and Health Informatics*	United States	2013	1656	27,850
10	*JMIR mHealth And uHealth*	Canada	2013	1544	16,156

Contributions of Authors and Institutes

A total of 114,841 authors (175,530 frequency) contributed to medical informatics from 2011 to 2020. As shown in Table 2, David W Bates (99 publications) was the most productive author, followed by Hua Xu (73 publications), and George Hripcsak (61 publications). Ian R White (5928 citations) was the most cited author, followed by David W Bates (2019 citations) and Joshua C Denny (1924 citations).

A total of 20,513 institutions contributed to the medical informatics field from 2011 to 2020. The number of institutions that issued more than 10 publications was 1385. Harvard University (1161 publications) was the most productive institution, followed by University of Toronto (503 publications), University of Washington (488 publications), and Columbia University (462 publications).

Table 2. The top 10 productive journals, authors, and institutions of medical informatics.

Rank	Author	Total publications	Total citations	Institution	Total publications	Total citations
1	David W Bates	99	2019	Harvard University	1161	19,471
2	Hua Xu	73	1691	University Toronto	503	10,904
3	George Hripcsak	61	1465	University Washington	488	6994
4	Dean F Sittig	61	1233	Columbia University	462	7744
5	Adam Wright	59	1102	University Michigan	455	6737
6	Hongfang Liu	58	873	Stanford University	422	7976
7	Joshua C Denny	51	1924	Vanderbilt University	389	8971
8	Chunhua Weng	51	818	University Penn	357	4084
9	Xiaoqian Jiang	48	565	Duke University	336	5375
10	Ian R White	47	5928	Mayo Clinic	326	5066

Co-authorship Analysis of Authors, Institutions, and Countries

Upon analyzing the 34,742 retrieved articles, there was an average of 5 co-authors for each article, revealing extensive co-authorships among authors in the field of medical informatics. We employed VOSviewer to analyze the co-authorship of authors, institutions, and countries/regions and then built the visualization network map.

We found that 304 productive authors published more than 15 articles. As shown in Figure 3, the largest collaborative network of productive authors comprising 234 authors was divided into 11 clusters of different colors. Hua Xu was the most active co-author with a total link strength of 162. The largest cluster (in red) involved 43 co-authors centering on Hua Xu, Xiaoqian Jiang, and Cui Tao. Figure 4 shows the collaborative network of 133 productive institutions that published more than 100 articles by 8 clusters of different colors. Harvard University was the most active co-author institution with a total link strength of 1484 and in the center of the green cluster. Figure 5 shows the largest collaborative network of countries/regions comprising 158 countries/regions divided into 4 different colored clusters. The United States was the most co-author country with a total link strength of 5495.

Figure 3. The collaborative network of productive authors participating in medical informatics publications from 2011 to 2020.

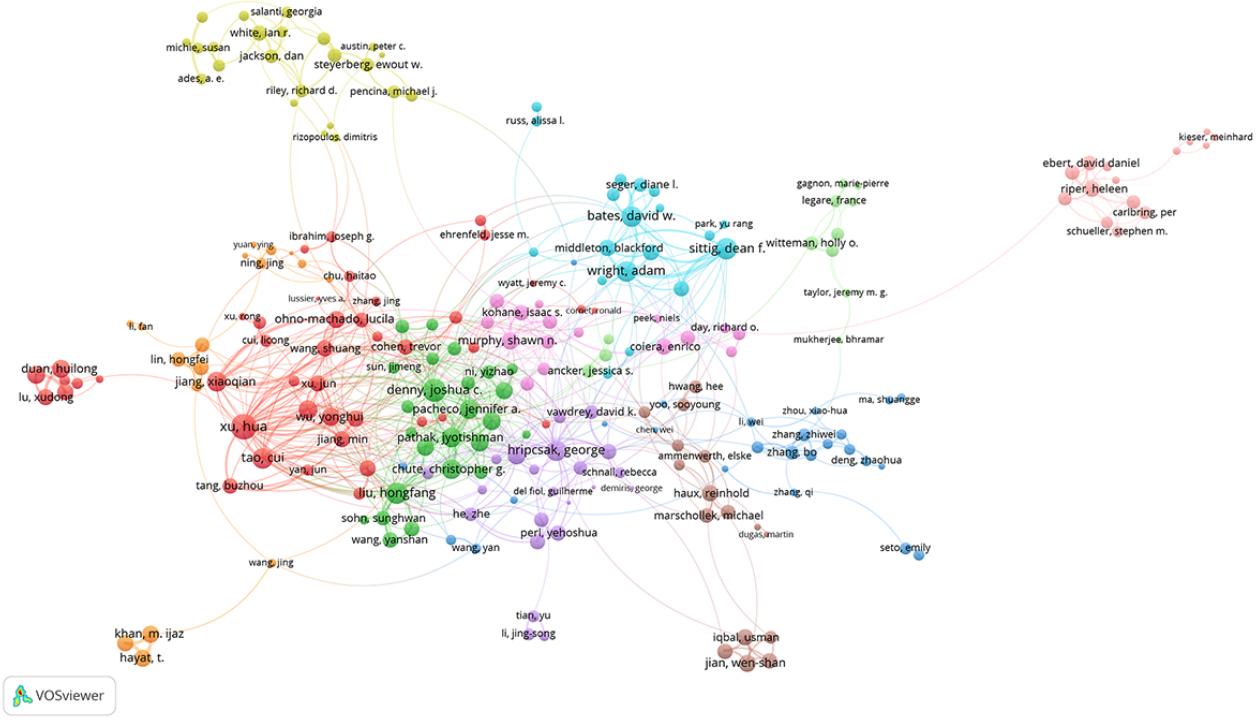


Figure 4. The collaborative network of productive institutions participating in medical informatics publications from 2011 to 2020.

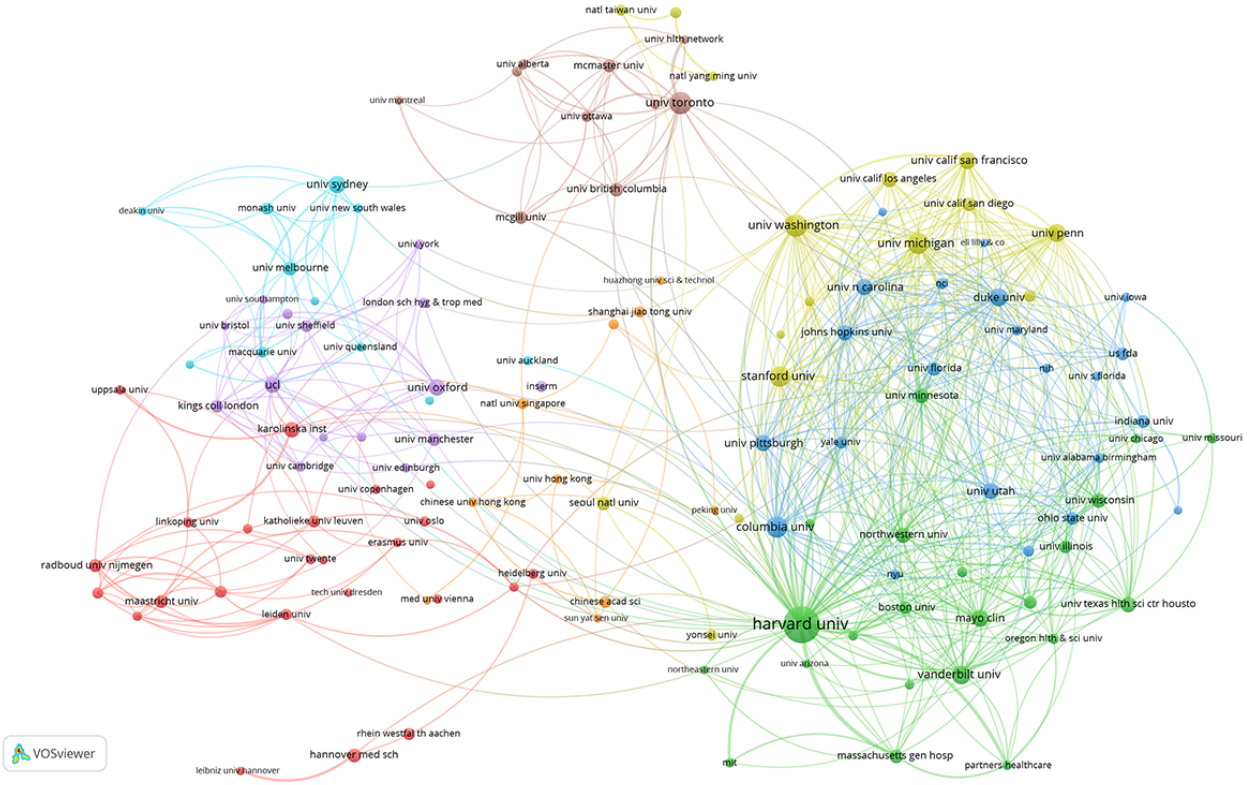
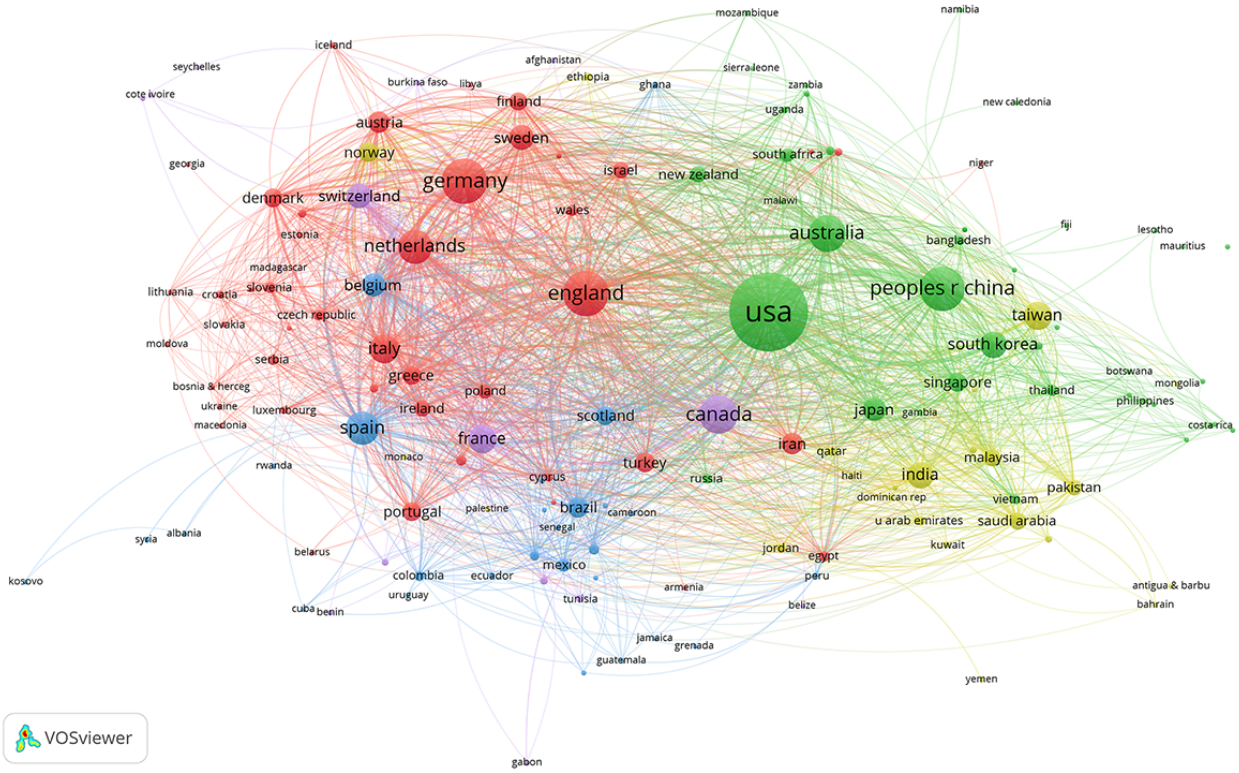


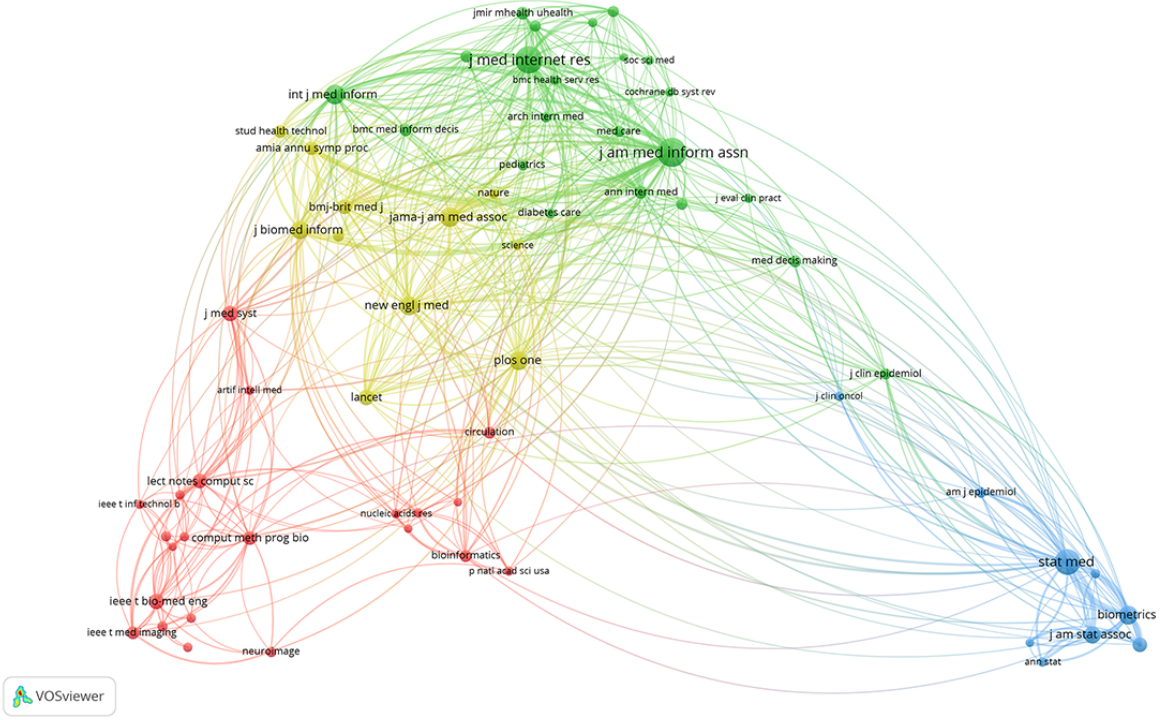
Figure 5. The collaborative network of countries/regions participating in medical informatics publications from 2011 to 2020.



Co-citation Analysis of References and Cited Journal

A total of 34,742 retrieved articles cited 700,628 references from 157,424 journals. The article published by Breiman in 2001 entitled “random forests” was the most cited reference. This article was cited 567 times in the retrieved publications on medical informatics and 40,253 times in WoSCC. Furthermore, the analysis of the distribution of cited journals helps identify the knowledge base of a certain field. The co-citation network of 64 cited journals with a minimum of 2000 citations was divided into 4 clusters of different colors. As shown in Figure 6, the red cluster centered on *IEEE Transactions on Biomedical Engineering*, and *IEEE Transactions on Medical Imaging*, *Computer Methods and Programs in Biomedicine*; the green cluster centered on *Journal of Medical Internet Research*, *Journal of the American Medical Informatics Association*, and *International Journal of Medical Informatics*; the blue cluster centered on *Statistics in Medicine*, *Biometrics*, *Journal of The American Statistical Association*; and the yellow cluster centered on *JAMA*, *New England Journal of Medicine*, *Lancet*, and *PLoS One*.

Figure 6. The co-citation network of highly cited journals on medical informatics from 2011 to 2020.



Co-occurrence Analysis of Author Keywords

The primary purpose of keywords is to provide fast access to scientific publications for researchers. In a bibliometric study, co-occurrence analysis of keywords effectively reflects the research hotspots in a scientific field [27,28]. This study analyzed the "author keywords" retrieved from WoSCC to represent the research hotspots. We employed VOSviewer to perform a co-occurrence analysis of the 143 high-frequency author keywords, which appeared more than 100 times from 2011 to 2020. The co-occurrence network map of high-frequency author keywords on medical informatics is shown in Figure 7. The most high-frequency author keyword was *EHRs* (with 1591 occurrences), followed by *mHealth* (n=1331), *machine learning* (n=994), *internet* (n=827), and *eHealth* (n=824). The 143 high-frequency author keywords formed 4 clusters: red, green, blue, and yellow. The red cluster is the largest one with 43 keywords regarding the research hotspots of AI in health care and medicine. The green cluster mainly focused on the research hotspots of mobile health; the blue cluster represented the research hotspots of implementation and evaluation of EHRs; the yellow cluster demonstrated the research hotspots of medical informatics technology application in public health.

We analyzed the theme evolution of the annual top 10 author keywords from 2011 to 2020, as shown in Figure 8. From 2011 to 2020, 28 author keywords entered the annual top 10 author keywords. The annual top 10 author keywords were constantly changing. *EHRs* was the only author keyword that has been in the annual top 10 for 10 consecutive years. *COVID-19*, which was ranked third in 2020, was the emerging theme of Medical Informatics.

Figure 7. The co-occurrence network of high-frequency author keywords on medical informatics publications from 2011 to 2020.

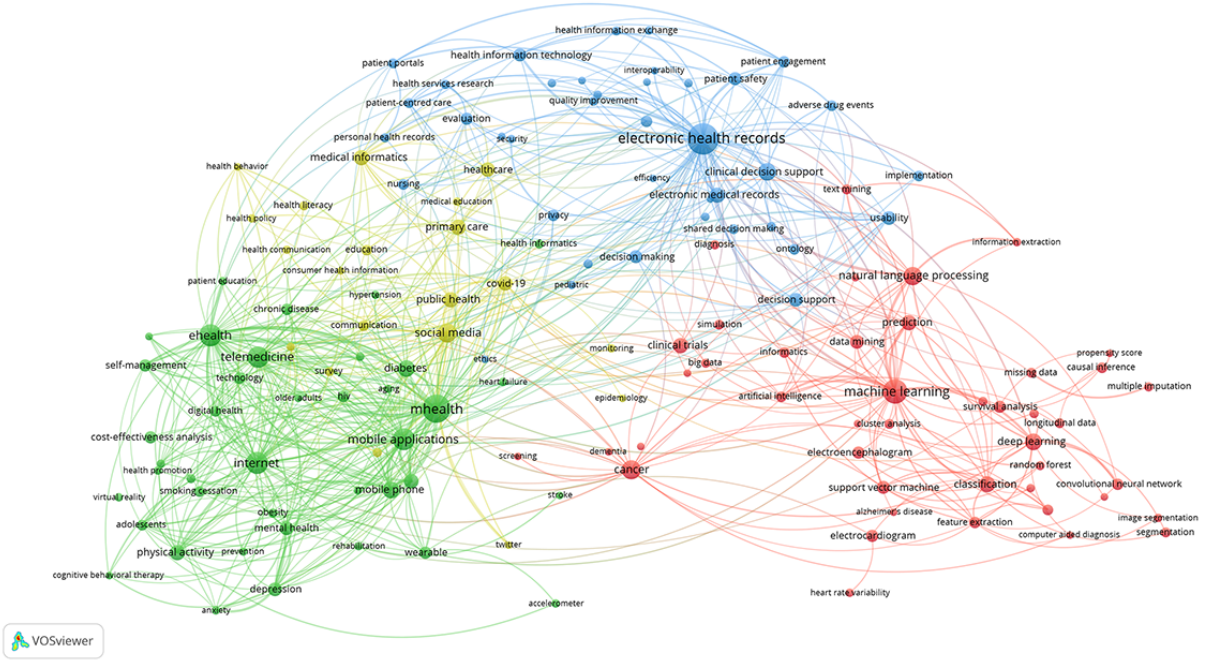
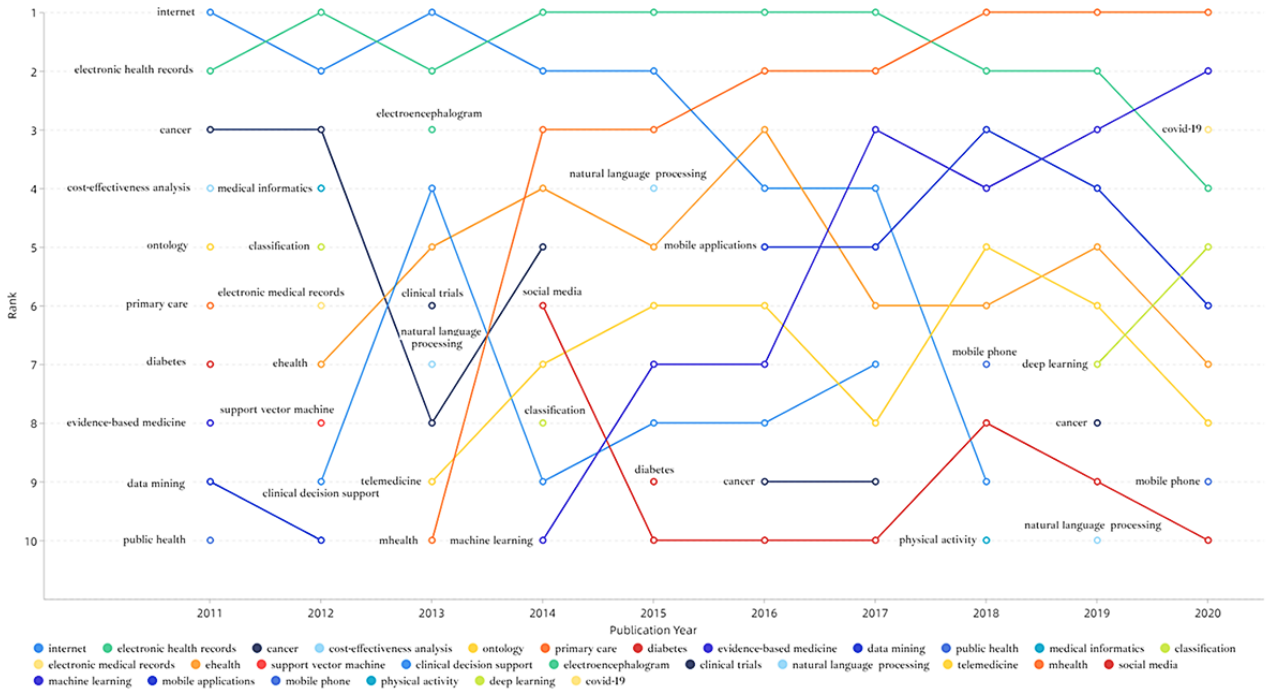


Figure 8. Theme evolution of the annual top 10 author keywords on medical informatics publications from 2011 to 2020.



Discussion

The Global Publication Trends of Medical Informatics

As shown in Multimedia Appendix 1, the global output of the entire life sciences and biomedical field shows a rapid growth trend from 2011 to 2020, except for 4 research areas. From 2011 to 2020, the increase rate of popularity of medical informatics far exceeded the growth rate of other research areas in life sciences and biomedical sciences, and it was almost 4 times the average increase rate, indicating the vast potential of medical informatics research in the future.

Figure 1 shows that most countries/regions have contributed to medical informatics research. Figure 2 shows that the United States was continuously ranked as the top productive country, indicating its outstanding contribution to the field of medical informatics. As the only developing and Asian country among the top 10 productive countries, China has leaped to second place in the top 10 productive countries in 2018, 2019, and 2020. The number of articles from China increased from 87 in

2011 to 946 in 2020, with a growth rate of 987.36%. China's growth rate was much higher than the average growth rate (193.86%, from 1987 in 2011 to 5839 in 2020), enabling it to improve its ranking rapidly. The rapid development of medical informatics in China may be attributed to the fact that China's medical reform in 2009 focused on the application of medical informatics technology. The number of articles from Germany increased from 124 in 2011 to 295 in 2020, with a growth rate of 137.90%. The growth rate of Germany was lower than the average growth rate (193.86%, from 1987 in 2011 to 5839 in 2020), which might be the main factor for its decline in ranking.

As shown in Table 1, the top 10 productive journals published 21,861 articles, accounting for 62.92% (21,861/34,742) of all medical informatics articles in the past 10 years. These journals have made substantial contributions to the development of medical informatics. From 2011 to 2020, the annual number of articles published in *Statistics in Medicine* was relatively stable, with minor fluctuations between 263 and 403. However, the ranking of *Statistics in Medicine* dropped from the first in 2011 to sixth in 2020, and its proportion dropped from 13.24% in 2011 to 5.53% in 2020. Except for 2014, the annual number of articles published by *Journal of Medical Internet Research* has continued to increase from 2011 to 2020. In particular, the annual number of articles published by *Journal of Medical Internet Research* has increased by 8.96 times, from 111 in 2011 to 1106 in 2020. *Journal of Medical Internet Research* was the most influential journal in medical informatics from 2011 to 2020 regardless of publications and citations. *IEEE Journal of Biomedical and Health Informatics* and *JMIR mHealth and uHealth*, published since 2013, were the fastest-growing journals in this field in the past 10 years.

The Global Collaboration Patterns of Medical Informatics

The international collaboration of authorship is attributed to a fast-growing increase in the number of outputs in a certain field [29]. In the past decade, an increasing number of authors, institutions, and countries/regions contributed to the productivity of medical informatics. From 2011 to 2020, more than 110,000 authors and 20,000 institutions from 161 countries/regions published medical informatics research. Additionally, to further demonstrate the dynamic changes of the authors in the field of medical informatics, we analyzed the authors who published medical informatics articles by categorizing the last decade as 2011-2015 and 2016-2020. Among the 79,829 authors who published medical informatics articles between 2016 and 2020, only 10,051 (12.59%) authors had previously published in this area. However, 69,778 (87.41%) authors published their first medical informatics papers during this period. Thus, many new researchers have flooded into medical informatics research in the past 5 years.

The top productive author David W Bates enjoyed an international reputation in medical informatics research, focusing on medical information technology to improve the safety and quality of medical care [30,31]. As shown in Figure 3, 70 of the 304 productive authors were not in the co-authorship network, indicating that the collaboration between productive authors still had certain limitations. Harvard University was always the top productive institution and at the center of the co-authorship network of institutions, indicating its substantial academic influences in medical informatics research. As shown in Figure 4, all the 133 productive institutions were in the co-authorship network, showing extensive collaborations between institutions worldwide. The United States continuously remained the top productive country and at the center of the co-authorship network of countries/regions, indicating its substantial academic influences in medical informatics research. As shown in Figure 5, 158 of the 161 countries/regions were in the co-authorship network, showing extensive collaborations between different countries.

The Basic Knowledge of Medical Informatics

Co-citation analysis can comprehensively demonstrate the knowledge base of a certain discipline [32]. As shown in Figure 6, the red cluster represents journals on computer science; the blue cluster represented journals on statistics science; the green cluster represented journals on medical informatics; the yellow cluster represented journals on general medicine and science and technology. In this study, co-citation analysis of cited journals showed that the knowledge base of medical informatics comes from medical informatics itself and disciplines such as computer science, general medicine, statistics, science and technology, and others.

The Research Hotspots of Medical Informatics

The co-occurrence analysis of high-frequency author keywords clarified the leading hotspots of medical informatics research. As shown in Figure 7, there are 4 main research hotspots on medical informatics from 2011 to 2020.

The red cluster focused on AI in health care and medicine. AI usually refers to computing technology that mimics or simulates the processes supported by human intelligence, which dramatically improves diagnosis and treatment accuracy and the entire clinical treatment process [33]. With the improvement in computer performance and the availability of big data from EHRs, the research and application of AI in health care and medicine have developed rapidly. With its advanced algorithms and learning capabilities, AI applications have helped medical professionals through symptom monitoring, predictive modeling, and decision support, especially in cancer and medical imaging [34,35].

The green cluster focused on mobile health. With the popularity of the internet and the rapid development of mobile communication devices and wearable devices, mobile health has been widely used in developed and developing countries [36]. Mobile health improves the ability of health systems to provide high-quality health care, especially in chronic disease, mental health, physical activity, HIV, and smoking cessation [37-40].

The blue cluster focused on the implementation and evaluation of EHRs. EHRs utilize information systems to store a digital format for patient and population health information [41]. Quantitative or qualitative methods were applied to evaluate the usability, interoperability, security, privacy, and other functions to improve EHRs continuously [42-44]. Health care professionals can access EHRs quickly and effectively to better

serve patients and the population, and these have great potential in improving medical efficiency and quality [45]. Implementing EHR and decision support helps clinicians make precise decisions to improve health care, reduce medical errors, and ensure patient safety [46-48].

The yellow cluster focused on medical informatics technology application in public health. The medical informatics technology represented by social media provides a series of possibilities for establishing multidirectional communication and interaction and quickly monitoring public emotions and activities. The application of new medical informatics technology can help increase the coverage and efficiency of public health services, especially in public communication, education, survey, engagement, and monitoring [49,50]. As our understanding of the most effective methods of using medical informatics technology to support public health research and practice matures, there will be more innovative applications of medical informatics technology in the field of public health, thereby making more remarkable contributions to improving population health.

The Theme Evolution and Emerging Frontiers of Medical Informatics

As shown in Figure 8, the content and ranking of the top 10 author keywords have evolved dramatically every year from 2011 to 2020. Only one of the top 10 author keywords in 2011 still appeared in the top 10 keywords in 2020. These were a microcosm of the rapid development of medical informatics and show that the theme of medical informatics research was significantly changing with the development of information technology.

EHRs was the only author keyword that continuously ranked in the top 10 during the past 10 years, and it was the most high-frequency keyword from 2011 to 2020. EHR was the most significant research hotspot of medical informatics throughout the past decade.

The *internet* consecutively ranked second from 2011 to 2015, but its ranking showed a gradual decline after 2016, showing that internet research based on traditional computers was the most concerned research theme in the early stages of medical informatics research in the past decade.

mHealth first appeared in the top 10 author keywords in 2013, and its ranking increased every year. Since 2018, mHealth is ranked as the number 1 author keyword for 3 consecutive years. Moreover, the author keywords *mobile applications* ranked sixth in 2020 and *mobile phones* ranked ninth in 2020, which were closely related to mHealth, showing that mHealth based on mobile devices has become the undisputed most prominent emerging theme in medical informatics.

Machine learning first appeared in the top 10 authors keywords in 2014 and has remained in the top 10 author keywords since then. The main methods of AI technology, machine learning and deep learning, were ranked second and fifth, respectively, in 2020, revealing that AI in health care was an emerging frontier of medical informatics. Especially, deep learning with the ability to mine a large amount of multimodal unstructured information and the ability to automate feature learning can promote the application of data-driven solutions in disease diagnosis and predicting prognosis [51,52].

The keywords related to health care and disease such as *cancer*, *diabetes*, *physical activity*, and *COVID-19* also appeared in the top 10 author keywords, indicating that the medical informatics technology has promising applications in treating, managing, monitoring, and preventing disease in the past decade. The outbreak of COVID-19 has had an unprecedented impact on global health, economy, and society. Various active response measures have been used to deal with the epidemic, and medical information also plays an important role, especially in coordinating medical resources, information dissemination, contact tracing, public education, and mental health intervention [53,54].

Limitations

Our research has some limitations. First, only English articles were retrieved in this study. Therefore, language bias may inevitably occur. Second, we did not evaluate the quality of publications, and some low-quality publications may have the same weight as high-quality publications. Finally, the data for this analysis were only extracted from WoSCC, excluding those from other databases such as Scopus, PubMed, or Google Scholar. Thus, publications appearing only through one of these databases may have been missed. Exploring ways to combine different data sources in future work is essential.

Conclusions

To our knowledge, this study provided the first comprehensive picture of global efforts on medical informatics in the past decade from a bibliometric analysis perspective. We summarize the recent advances in medical informatics in the past decade and shed light on their publication trends, influential journals, global collaboration patterns, basic knowledge, research hotspots, theme evolution, and emerging frontiers. These findings will accurately and quickly grasp the research trends and provide valuable guidance for future medical informatics research.

Acknowledgments

This work was funded by the Hwa Mei fund (Grant No. 2019HMKY23), Scientific Projects of the Education Department of Liaoning Province (Grant No. LJKR0273), Scientific Projects of the Education Department of Liaoning Province (Grant No. ZF2019026).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The global output of the entire life sciences and biomedical field.
[XLSX File (Microsoft Excel File), 38 KB - medinform_v10i4e33842_app1.xlsx]

References

1. Hasman A, Haux R, Albert A. A systematic view on medical informatics. Computer Methods and Programs in Biomedicine 1996 Nov;51(3):131-139. [doi: 10.1016/s0169-2607(96)01769-5]
2. Masic I. Five periods in development of medical informatics. Acta Inform Med 2014 Feb;22(1):44-48 [FREE Full text] [doi: 10.5455/aim.2014.22.44-48] [Medline: 24648619]
3. Lorenzi NM, Gardner RM, Pryor TA, Stead WW. Medical informatics: the key to an organization's place in the new health care environment. J Am Med Inform Assoc 1995;2(6):391-392 [FREE Full text] [doi: 10.1136/jamia.1995.96157832] [Medline: 8581555]
4. Deng H, Wang J, Liu X, Liu B, Lei J. Evaluating the outcomes of medical informatics development as a discipline in China: A publication perspective. Comput Methods Programs Biomed 2018 Oct;164:75-85. [doi: 10.1016/j.cmpb.2018.07.001] [Medline: 30195433]
5. Imhoff M, Webb A, Goldschmidt A, European Society of Intensive Care Medicine. ESCIM. Health informatics. Intensive Care Med 2001 Jan 19;27(1):179-186. [doi: 10.1007/pl00020869] [Medline: 11280631]
6. Fridsma DB. Health informatics: a required skill for 21st century clinicians. BMJ 2018 Jul 12;362:k3043. [doi: 10.1136/bmj.k3043] [Medline: 30002063]
7. Stead WW, Lorenzi NM. Health informatics: linking investment to value. J Am Med Inform Assoc 1999 Sep 01;6(5):341-348 [FREE Full text] [doi: 10.1136/jamia.1999.0060341] [Medline: 10495093]
8. Tremblay MC, Deckard GJ, Klein R. Health informatics and analytics - building a program to integrate business analytics across clinical and administrative disciplines. J Am Med Inform Assoc 2016 Jul;23(4):824-828. [doi: 10.1093/jamia/ocw055] [Medline: 27274022]
9. Young H, Belanger T. Young, T. In: The ALA Glossary of Library and Information Science. Chicago, IL: American Library Association; 1983.
10. Bansard JY, Rebholz-Schuhmann D, Cameron G, Clark D, van Mulligen E, Beltrame E, et al. Medical informatics and bioinformatics: a bibliometric study. IEEE Trans Inf Technol Biomed 2007 May;11(3):237-243 [FREE Full text] [doi: 10.1109/titb.2007.894795] [Medline: 17521073]
11. Cao J, Lim Y, Sengoku S, Guo X, Kodama K. Exploring the Shift in International Trends in Mobile Health Research From 2000 to 2020: Bibliometric Analysis. JMIR Mhealth Uhealth 2021 Sep 08;9(9):e31097 [FREE Full text] [doi: 10.2196/31097] [Medline: 34494968]
12. Blanc X, Collet TH, Auer R, Fischer R, Locatelli I, Iriarte P, et al. Publication trends of shared decision making in 15 high impact medical journals: a full-text review with bibliometric analysis. BMC Med Inform Decis Mak 2014 Aug 09;14:71 [FREE Full text] [doi: 10.1186/1472-6947-14-71] [Medline: 25106844]
13. Armfield NR, Edirippulige S, Caffery LJ, Bradford NK, Grey JW, Smith AC. Telemedicine--a bibliometric and content analysis of 17,932 publication records. Int J Med Inform 2014 Oct;83(10):715-725. [doi: 10.1016/j.ijmedinf.2014.07.001] [Medline: 25066950]
14. Yang YT, Iqbal U, Ching JHY, Ting JBS, Chiu HT, Tamashiro H, et al. Trends in the growth of literature of telemedicine: A bibliometric analysis. Comput Methods Programs Biomed 2015 Dec;122(3):471-479. [doi: 10.1016/j.cmpb.2015.09.008] [Medline: 26415760]
15. Waqas A, Teoh SH, Lapão LV, Messina LA, Correia JC. Harnessing Telemedicine for the Provision of Health Care: Bibliometric and Scientometric Analysis. J Med Internet Res 2020 Oct 02;22(10):e18835 [FREE Full text] [doi: 10.2196/18835] [Medline: 33006571]
16. Takahashi R, Kajikawa Y. Computer-aided diagnosis: A survey with bibliometric analysis. Int J Med Inform 2017 May;101:58-67. [doi: 10.1016/j.ijmedinf.2017.02.004] [Medline: 28347448]
17. Chen X, Xie H, Wang FL, Liu Z, Xu J, Hao T. A bibliometric analysis of natural language processing in medical research. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):14 [FREE Full text] [doi: 10.1186/s12911-018-0594-x] [Medline: 29589569]
18. Guo Y, Hao Z, Zhao S, Gong J, Yang F. Artificial Intelligence in Health Care: Bibliometric Analysis. J Med Internet Res 2020 Jul 29;22(7):e18228 [FREE Full text] [doi: 10.2196/18228] [Medline: 32723713]
19. Taj F, Klein MCA, van Halteren A. Digital Health Behavior Change Technology: Bibliometric and Scoping Review of Two Decades of Research. JMIR Mhealth Uhealth 2019 Dec 13;7(12):e13311 [FREE Full text] [doi: 10.2196/13311] [Medline: 31833836]
20. Fornazin MP, Bruno EP, Costa de Castro L, de Castro Silva SLF. From medical informatics to digital health: a bibliometric analysis of the research field. In: AMCIS 2021 Proceedings. 2021 Presented at: AMCIS 2021; August 9-13, 2021; Virtual URL: https://aisel.aisnet.org/amcis2021/healthcare_it/sig_health/18

21. Agarwal A, Durairajanayagam D, Tatagari S, Esteves SC, Harlev A, Henkel R, et al. Bibliometrics: tracking research impact by selecting the appropriate metrics. Asian J Androl 2016;18(2):296-309 [FREE Full text] [doi: 10.4103/1008-682X.171582] [Medline: 26806079]
22. Newman MEJ. Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci U S A 2004 Apr 06;101 Suppl 1:5200-5205 [FREE Full text] [doi: 10.1073/pnas.0307545100] [Medline: 14745042]
23. Romero L, Portillo-Salido E. Trends in Sigma-1 Receptor Research: A 25-Year Bibliometric Analysis. Front Pharmacol 2019;10:564 [FREE Full text] [doi: 10.3389/fphar.2019.00564] [Medline: 31178733]
24. VOSviewer. URL: http://www.vosviewer.com [accessed 2022-03-29]
25. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010 Aug;84(2):523-538 [FREE Full text] [doi: 10.1007/s11192-009-0146-3] [Medline: 20585380]
26. Chen C. Searching for intellectual turning points: progressive knowledge domain visualization. Proc Natl Acad Sci U S A 2004 Apr 06;101 Suppl 1:5303-5310 [FREE Full text] [doi: 10.1073/pnas.0307513100] [Medline: 14724295]
27. Zhao D, Li J, Seehus C, Huang X, Zhao M, Zhang S, et al. Bibliometric analysis of recent sodium channel research. Channels (Austin) 2018;12(1):311-325 [FREE Full text] [doi: 10.1080/19336950.2018.1511513] [Medline: 30134757]
28. Zhou H, Tan W, Qiu Z, Song Y, Gao S. A bibliometric analysis in gene research of myocardial infarction from 2001 to 2015. PeerJ 2018;6:e4354 [FREE Full text] [doi: 10.7717/peerj.4354] [Medline: 29456889]
29. Corrales-Reyes I. Co-authorship and scientific collaboration networks in Medwave. Medwave 2017 Dec 28;17(09):e7103-e7103. [doi: 10.5867/medwave.2017.09.7103]
30. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc 2014;21(3):487-491. [doi: 10.1136/amiajnl-2013-001813] [Medline: 24166725]
31. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. J Am Med Inform Assoc 2013 May 01;20(3):489-493 [FREE Full text] [doi: 10.1136/amiajnl-2012-001089] [Medline: 23011124]
32. Zou LX, Sun L. Global diabetic kidney disease research from 2000 to 2017: A bibliometric analysis. Medicine (Baltimore) 2019 Feb;98(6):e14394 [FREE Full text] [doi: 10.1097/MD.0000000000014394] [Medline: 30732183]
33. Tran BX, Vu GT, Ha GH, Vuong QH, Ho MT, Vuong TT, et al. Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. J Clin Med 2019 Mar 14;8(3):360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
34. Hirasawa T, Aoyama K, Tanimoto T, Ishihara S, Shichijo S, Ozawa T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018 Jul;21(4):653-660. [doi: 10.1007/s10120-018-0793-2] [Medline: 29335825]
35. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol 2019 Mar;20(3):405-410 [FREE Full text] [doi: 10.3348/kjr.2019.0025] [Medline: 30799571]
36. Silva BMC, Rodrigues JJPC, de la Torre Díez I, López-Coronado M, Saleem K. Mobile-health: A review of current state in 2015. J Biomed Inform 2015 Aug;56:265-272 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.003] [Medline: 26071682]
37. Miller AS, Cafazzo JA, Seto E. A game plan: Gamification design principles in mHealth applications for chronic disease management. Health Informatics J 2016 Jun;22(2):184-193 [FREE Full text] [doi: 10.1177/1460458214537511] [Medline: 24986104]
38. Berry K, Salter A, Morris R, James S, Bucci S. Assessing Therapeutic Alliance in the Context of mHealth Interventions for Mental Health Problems: Development of the Mobile Agnew Relationship Measure (mARM) Questionnaire. J Med Internet Res 2018 Apr 19;20(4):e90 [FREE Full text] [doi: 10.2196/jmir.8252] [Medline: 29674307]
39. Guo Y, Xu Z, Qiao J, Hong YA, Zhang H, Zeng C, et al. Development and Feasibility Testing of an mHealth (Text Message and WeChat) Intervention to Improve the Medication Adherence and Quality of Life of People Living with HIV in China: Pilot Randomized Controlled Trial. JMIR Mhealth Uhealth 2018 Sep 04;6(9):e10274 [FREE Full text] [doi: 10.2196/10274] [Medline: 30181109]
40. Bustamante LA, Gill Ménard C, Julien S, Romo L. Behavior Change Techniques in Popular Mobile Apps for Smoking Cessation in France: Content Analysis. JMIR Mhealth Uhealth 2021 May 13;9(5):e26082 [FREE Full text] [doi: 10.2196/26082] [Medline: 33983130]
41. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. J Med Internet Res 2005 Mar 14;7(1):e3 [FREE Full text] [doi: 10.2196/jmir.7.1.e3] [Medline: 15829475]
42. Horsky J, McColgan K, Pang JE, Melnikas AJ, Linder JA, Schnipper JL, et al. Complementary methods of system usability evaluation: surveys and observations during software design and development cycles. J Biomed Inform 2010 Oct;43(5):782-790 [FREE Full text] [doi: 10.1016/j.jbi.2010.05.010] [Medline: 20546936]
43. Praveen D, Patel A, Raghu A, Clifford GD, Maulik PK, Mohammad Abdul A, et al. SMARTHealth India: Development and Field Evaluation of a Mobile Clinical Decision Support System for Cardiovascular Diseases in Rural India. JMIR Mhealth Uhealth 2014 Dec 08;2(4):e54 [FREE Full text] [doi: 10.2196/mhealth.3568] [Medline: 25487047]

44. Fossum M, Ehnfors M, Fruhling A, Ehrenberg A. An evaluation of the usability of a computerized decision support system for nursing homes. Appl Clin Inform 2011;2(4):420-436 [FREE Full text] [doi: 10.4338/ACI-2011-07-RA-0043] [Medline: 23616886]
45. Opipari-Arrigan L, Dykes DMH, Saeed SA, Thakkar S, Burns L, Chini BA, et al. Technology-Enabled Health Care Collaboration in Pediatric Chronic Illness: Pre-Post Interventional Study for Feasibility, Acceptability, and Clinical Impact of an Electronic Health Record-Linked Platform for Patient-Clinician Partnership. JMIR Mhealth Uhealth 2020 Nov 26;8(11):e11968. [doi: 10.2196/11968] [Medline: 33242014]
46. Payne TH, Hines LE, Chan RC, Hartman S, Kapusnik-Uner J, Russ AL, et al. Recommendations to improve the usability of drug-drug interaction clinical decision support alerts. J Am Med Inform Assoc 2015 Nov;22(6):1243-1250. [doi: 10.1093/jamia/ocv011] [Medline: 25829460]
47. Galanter WL, Hier DB, Jao C, Sarne D. Computerized physician order entry of medications and clinical decision support can improve problem list documentation compliance. Int J Med Inform 2010 May;79(5):332-338. [doi: 10.1016/j.ijmedinf.2008.05.005] [Medline: 18599342]
48. Miller A, Moon B, Anders S, Walden R, Brown S, Montella D. Integrating computerized clinical decision support systems into clinical work: A meta-synthesis of qualitative research. Int J Med Inform 2015 Dec;84(12):1009-1018. [doi: 10.1016/j.ijmedinf.2015.09.005] [Medline: 26391601]
49. Capurro D, Cole K, Echavarría MI, Joe J, Neogi T, Turner AM. The use of social networking sites for public health practice and research: a systematic review. J Med Internet Res 2014 Mar 14;16(3):e79 [FREE Full text] [doi: 10.2196/jmir.2679] [Medline: 24642014]
50. Zhang Y, Cao B, Wang Y, Peng TQ, Wang X. When Public Health Research Meets Social Media: Knowledge Mapping From 2000 to 2018. J Med Internet Res 2020 Aug 13;22(8):e17582 [FREE Full text] [doi: 10.2196/17582] [Medline: 32788156]
51. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan;25(1):24-29 [FREE Full text] [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
52. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/JBHI.2016.2636665] [Medline: 28055930]
53. Bao H, Cao B, Xiong Y, Tang W. Digital Media's Role in the COVID-19 Pandemic. JMIR Mhealth Uhealth 2020 Sep 18;8(9):e20156 [FREE Full text] [doi: 10.2196/20156] [Medline: 32530817]
54. Crawford A, Serhal E. Digital Health Equity and COVID-19: The Innovation Curve Cannot Reinforce the Social Gradient of Health. J Med Internet Res 2020 Jun 02;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]

Abbreviations

AI: artificial intelligence
EHRs: electronic health records
WoSCC: Web of Science Core Collection

Edited by C Lovis; submitted 16.10.21; peer-reviewed by J Wang, JK Kumar; comments to author 10.11.21; revised version received 26.12.21; accepted 31.01.22; published 21.04.22.

Please cite as:
He X, Peng C, Xu Y, Zhang Y, Wang Z
Global Scientific Research Landscape on Medical Informatics From 2011 to 2020: Bibliometric Analysis
JMIR Med Inform 2022;10(4):e33842
URL: https://medinform.jmir.org/2022/4/e33842
doi:10.2196/33842
PMID:35451986

Original Paper

The Effectiveness of the Capacity Building and Mentorship Program in Improving Evidence-Based Decision-making in the Amhara Region, Northwest Ethiopia: Difference-in-Differences Study

Moges Asressie Chanyalew[1], MSc; Mezgebu Yitayal[2], PhD; Asmamaw Atnafu[2], PhD; Shegaw Anagaw Mengiste[3], PhD; Binyam Tilahun[1], PhD

[1]Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[2]Department of Health Systems and Policy, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[3]Management Information Systems, School of Business, University of South-Eastern Norway, Notodden, Norway

Corresponding Author:
Moges Asressie Chanyalew, MSc
Department of Health Informatics, Institute of Public Health
College of Medicine and Health Sciences
University of Gondar
Gondar
Ethiopia
Phone: 251 911617734
Fax: 251 5882221626
Email: mogesabu@gmail.com

Abstract

Background: Weak health information systems (HISs) hobble countries' abilities to effectively manage and distribute their resources to match the burden of disease. The Capacity Building and Mentorship Program (CBMP) was implemented in select districts of the Amhara region of Ethiopia to improve HIS performance; however, evidence about the effectiveness of the intervention was meager.

Objective: This study aimed to determine the effectiveness of routine health information use for evidence-based decision-making among health facility and department heads in the Amhara region, Northwest Ethiopia.

Methods: The study was conducted in 10 districts of the Amhara region: five were in the intervention group and five were in the comparison group. We employed a quasi-experimental study design in the form of a pretest-posttest comparison group. Data were collected from June to July 2020 from the heads of departments and facilities in 36 intervention and 43 comparison facilities. The sample size was calculated using the double population formula, and we recruited 172 participants from each group. We applied a difference-in-differences analysis approach to determine the effectiveness of the intervention. Heterogeneity of program effect among subgroups was assessed using a triple differences method (ie, difference-in-difference-in-differences [DIDID] method). Thus, the β coefficients, 95% CIs, and *P* values were calculated for each parameter, and we determined that the program was effective if the interaction term was significant at $P<.05$.

Results: Data were collected using the endpoint survey from 155 out of 172 (90.1%) participants in the intervention group and 166 out of 172 (96.5%) participants in the comparison group. The average level of information use for the comparison group was 37.3% (95% CI 31.1%-43.6%) at baseline and 43.7% (95% CI 37.9%-49.5%) at study endpoint. The average level of information use for the intervention group was 52.2% (95% CI 46.2%-58.3%) at baseline and 75.8% (95% CI 71.6%-80.0%) at study endpoint. The study indicated that the net program change over time was 17% (95% CI 5%-28%; $P=.003$). The subgroup analysis also indicated that location showed significant program effect heterogeneity, with a DIDID estimate equal to 0.16 (95% CI 0.026-0.29; $P=.02$). However, sex, age, educational level, salary, and experience did not show significant heterogeneity in program effect, with DIDID estimates of 0.046 (95% CI –0.089 to 0.182), –0.002 (95% CI –0.015 to 0.009), –0.055 (95% CI –0.190 to 0.079), –1.63 (95% CI –5.22 to 1.95), and –0.006 (95% CI –0.017 to 0.005), respectively.

Conclusions: The CBMP was effective at enhancing the capacity of study participants in using the routine HIS for decision-making. We noted that urban facilities had benefited more than their counterparts. The intervention has been shown to

produce positive outcomes and should be scaled up to be used in other districts. Moreover, the mentorship modalities for rural facilities should be redesigned to maximize the benefits.

Trial Registration: Pan African Clinical Trials Registry PACTR202001559723931; https://tinyurl.com/3j7e5ka5

(JMIR Med Inform 2022;10(4):e30518) doi:10.2196/30518

KEYWORDS

capacity building; mentorship; mentoring; mentor; training; data use; information use; facility head; department head; quasi-experiment; difference-in-differences; Ethiopia; Amhara; weak health information system; HIS; health information system; CBMP; DID; decision-making; Africa; evidence based; effectiveness

Introduction

A health information system (HIS) is an intersection between health care business processes and information systems to deliver better health care services [1]. An effective and integrated HIS is the foundation of a strong health system and provides underpinnings for decision-making [2,3]. It has much to offer in managing health care costs and improving health care quality [4,5]. Effective decision-making to improve public health care essentially depends on the availability of reliable data [6].

Countries have made tremendous efforts to enhance data use practice for patient care and management. For example, a granular ontology model for maternal and child HISs and a national acute care information platform were implemented and improved data analysis skills and policy making in Pakistan [6] and Sri Lanka [7], respectively. On the other hand, timely feedback on health system performance was implemented in sub-Saharan African countries and resulted in enhanced decision-making among leaders [8]. Likewise, a data-driven quality improvement intervention in Mozambique, Rwanda, and Zambia [9], as well as a data use workshop in Zanzibar and the United Republic of Tanzania [10], were implemented and brought a shift from a lack of awareness to collaborative ownership and improved local use of target indicators to drive change, respectively.

In Botswana, a task-shifting initiative (ie, development of a dedicated monitoring and evaluation cadre) was implemented to strengthen monitoring and evaluation and build a sustainable HIS. As a result, the intervention brought increased use of health data for disease surveillance, operational research, and planning purposes [11]. The Feedback and Analytic Comparison Tool in Egypt [12] and the District Health Profile tool in Kenya [13] were implemented as change drivers; they helped health workers to identify gaps and facilitated data-informed decision-making. Moreover, a partnership-mentoring model implemented in the public hospitals of Ethiopia resulted in a 60% improvement of management indicators [14].

Previous work has indicated that training, supervision, a good perceived culture of information use, having standard indicators, competence on routine HIS (RHIS) tasks, technology enhancement along with capacity building activities, and feedback systems were positively associated with routine health information use [15-18]. Despite this, there has been a concern when using this information in strategic decision-making among health workers [19] as well as a concern that the RHIS was unfairly used to enhance evidence-based decision-making [15]. According to Mate et al [20], incomplete, inaccurate, and untimely data have been challenges of data use. Hoxha et al [16] also documented that the technical, organizational, and behavioral attributes of RHIS data remain challenges in health data use.

The investment in health systems infrastructure or training for clinicians and administrators in low- and middle-income countries (LMIC) has been low [21]. Weak HISs hobble the ability of many LMIC to distribute their resources to match the burden of disease [22], and lack of data use is disempowering staff and those seeking to support them from making progress in setting-relevant research and quality improvement [7]. Considering the low level of health data use among health workers, the Amhara Regional Health Bureau (ARHB), in collaboration with the University of Gondar (UoG) in Ethiopia, introduced the Capacity Building and Mentorship Program (CBMP) in 2019. The initiative has been implemented in five selected districts of the region since then. However, information was meager about the effectiveness of the intervention on the use of information in the study area. Therefore, this study aimed to determine the level of health information use among health facilities and department heads in the Amhara region, Northwest Ethiopia.

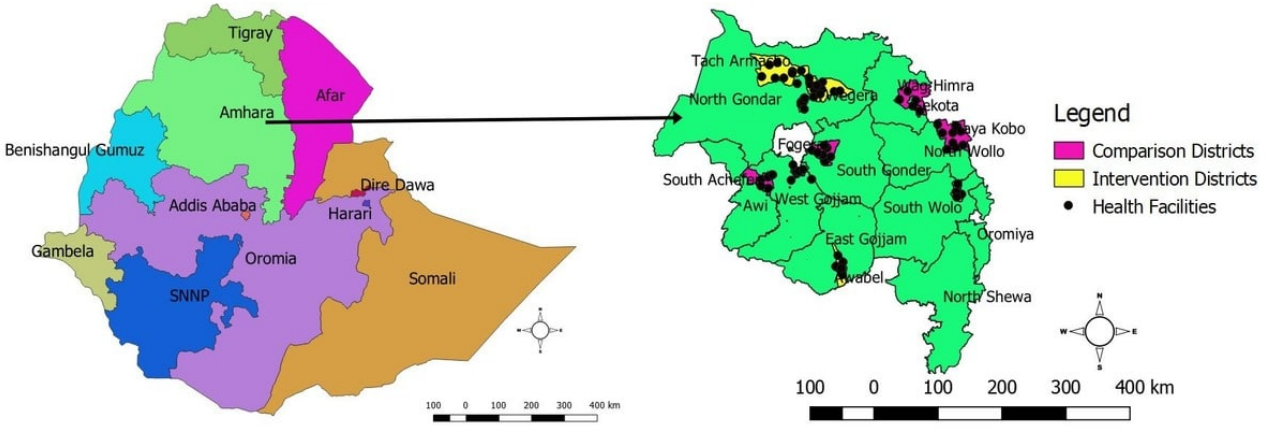
Methods

Study Setting and Design

Baseline data were collected from April to May 2019 in 10 selected districts of the Amhara region: five were in the intervention group and five were in the comparison group. A year later, endpoint data were collected from June to July 2020 in the same districts. Thus, to estimate the effectiveness of the intervention, a quasi-experimental, nonequivalent, control group study design was employed [23].

The Amhara region is located between 8˚45′ N and 13˚45′ N latitude and 35˚46′ E and 40˚25′ E longitude in Northwest Ethiopia [24,25]. It is subdivided administratively into 12 zones and three town administrations (Figure 1). The zones and the town administrations are again subdivided into a total of 189 districts, of which 39 are urban towns. As of 2020, the total population of the region was 22,292,890. The ratio of males to females was close to 1. The region has been implementing a three-tier health system comprised of primary, secondary, and tertiary levels. There are nine referral hospitals, 71 primary hospitals, 954 health centers, and 3450 health posts that are providing health services in the region [26].

Figure 1. Map of the study area in the Amhara Region, Northwest Ethiopia, 2021.



Participants

The source population of this study included all health department and health facility heads. Individuals responsible for the health departments or health institutions that were expected to use routine health data and who were at the selected facilities for at least 6 months were included in the study. However, health department heads who were on leave, retired, and not supposed to use routine health information for patient monitoring and follow-up were excluded from the study. Since it was an intervention effectiveness study, newly inaugurated facilities with patient stays of up to 3 months were also excluded from the study.

Interventions

The CBMP is an innovative approach that has been implemented since 2019 in selected districts of Ethiopia through the joint venture between the Federal Ministry of Health (FMoH) and selected universities; its goal is to strengthen the national HIS through proper data documentation, information use, and digitalization. The ARHB along with the UoG were responsible for implementing the intervention in the Amhara region. The intervention had two components: tailored training and mentorship. It targeted service delivery unit heads, case team leaders, health department heads, health facility heads, program officers, and district office managers [27].

The intervention was designed primarily to improve the capacity of health workers at different levels. Thus, data quality and information use training manuals were distributed to health facilities, and participants from intervention districts received training on DHIS2 (District Health Information Software 2) data analytics, visualization, presentation, troubleshooting, and data use for decision-making. Furthermore, HIS resources (ie, registers, tally sheets, and computers) were provided to health facilities to enhance the availability and quality of data, and joint review meetings, which served as a platform for discussion, were organized every 6 months [28].

Mentoring is a strategic development activity that supports health workers to attain the vision and goals of the organization [29]. The UoG recruited mentors from departments such as health informatics, health systems and policy, epidemiology and biostatistics, and health education. Thus, training was organized and provided to mentors to introduce them to the HIS strategies being implemented nationally and to provide them with mentorship skills. The trained mentors conducted mentorship programs at each health facility department every quarter for 1 year using the mentorship checklist. As a result, four rounds of mentorship programs were conducted. After that, the mentors with their mentees developed action plans that indicated activities to be accomplished, responsible persons, and implementation period. Based on the mentorship findings, mentees provided detailed written feedback that contained the strengths, weaknesses, and next steps in the performance of the RHIS.

Outcomes

The dependent variable of the study was the level of information use. The concept of *information used for action in the health care system* was applied. The practice of routine health data use is a process that encompasses gap identification, prioritization, root cause analysis, action plan development, and follow-up. Thus, we used a composite indicator to calculate the average value of information use. The five components of the outcome variable were as follows: identifying indicators in the department, calculating targets versus achievements, providing feedback to health workers at the lower levels, calculating program coverage, and evidence showing the use of data to inform decisions. We calculated the average value of these five indicators and compared the level of information use among intervention and comparison groups [15,30].

Sample Size

The study employed a quasi-experimental design with pre- and postassessment and intervention and comparison groups. Considering the difference in the level of information use among intervention and comparison districts after the intervention, we employed a double population proportion formula to estimate the sample size. Mathematically, the sample size was determined using the following formula [31]:

$$N = (K [P_1 (1 - P_1) + P_2 (1 - P_2)]) / ([P_1 - P_2]^2)$$

where N is the sample size; P_1 is the anticipated proportion of facilities with the attribute of interest (ie, level of information use after the intervention, assuming a 15% increase in information use and considered as 84%) [32]; P_2 is the proportion of data use with no intervention, taken as 69%; K is the constant at an α value of .05 and a β value of .2, taken as

7.9; and power (1 – β) was 80% [33,34]. With these assumptions, the sample size was calculated to be 122 units per department for each group. Though the source population was finite (<10,000), the sample size was corrected using the correction formula. Considering a design effect of 1.5 and a 5% nonresponse rate, the final sample size was 172 for each group. Thus, the overall sample size for both groups combined was 344 individuals.

As a quasi-experimental study, five districts were recruited for each arm (ie, intervention and comparison groups). The intervention districts were selected by the FMoH from among the low-performance districts regarding RHIS activities in the region. However, comparison districts were chosen randomly among the 146 districts in the region. We applied a multistage sampling procedure to select the study participants and developed a sampling frame that contained a list of the heads of departments and facilities in the selected districts. Thus, study participants were selected from the study population using a simple random sampling technique.

Data Collection Tools and Procedures

Data collection tools were developed based on the Performance of Routine Information System Management tools (version 3) and adapted to the local context [35]. In this paper, we applied the Information Use Assessment Tool and Organizational and Behavioral Assessment Tool (OBAT). The tools were piloted in two districts, Injibara and Debre Tabor, for validity and reliability checks; the districts were out of the study area but comparable with the study sites. The reliability assessment score showed a Cronbach α of .92 for the Likert scale, which indicated that the tool was consistent in measuring the outcome of interest [36].

The Information Use Assessment Tool is an interviewer-administered tool. It was used to examine the health facilities' report production, information display, discussions, and decisions based on the RHIS, planning, supervision, and mentorship. The OBAT is a self-administered tool that is used to identify information about the technical, organizational, and behavioral constraints for routine health data use. Eight data collectors and two supervisors participated in the data collection. The principal investigator (PI) delivered training on data recording, document review, and ethical consideration to data collectors and supervisors for 2 days. A data quality checklist was developed and applied during data collection to maintain the quality of data. Daily feedback was provided to data collectors by supervisors. The PI led and coordinated the overall data collection process.

Data collectors requested permission from the facility heads to access the documents and departments. In addition, they provided information about the purpose of the study and obtained written consent from selected participants before interviewing them. Following that, they prepared participants for the interview. They reviewed source documents and charts posted in the department or unit, observed the discussion points made among members of the management body in the logbook, and collected data using the study tool. Subsequently, respondents were provided with a self-administered tool (ie, the OBAT) and informed to take an ample amount of time to complete the questionnaire.

Assignment Methods

The FMoH with the ARHB selected five intervention districts based on predefined criteria. All of these districts were low performers regarding the RHIS activities. They had a low level of information use and poor data quality but could potentially improve their performance if given the intervention. As a result, random assignments of districts to comparison and intervention arms were not applicable. However, the research team, in collaboration with the ARHB, selected five comparison districts to help in measuring the effectiveness of the intervention. Therefore, the intervention groups received usual service and the new CBMP intervention (ie, tailored training and mentorship), and the comparison groups received usual service (ie, supervision and review meeting by routine service).

Blinding

The nature of the intervention was designed to be implemented by mobile mentors. As a result, we were unable to mask health workers and program implementers. However, all study participants and data collectors were blinded to the research question and hypothesis that the team generated during implementation, baseline, and endpoint data collection.

Statistical Methods

The team scrutinized the data to identify missing values before entering the data into the software. The data entry template was developed using EpiData software (version 3.1; EpiData Association) by applying the commands and skipping patterns that minimized errors during entry. Thus, cleaned data were entered into EpiData and exported to R software (version 4.0.4; The R Foundation) in CSV file format to compute the effect size of the intervention. R software has different built-in statistical packages that enable researchers to run statistical models and test the hypothesis in question. Descriptive statistics, such as mean and percentage, were calculated. Tables and graphs were also used for presenting findings.

The data were collected from the intervention and comparison districts before and after the implementation of the intervention. Thus, it entailed the difference-in-differences (DID) method to determine the effectiveness of the CBMP. As alluded to by different scholars, the DID method is one of the most frequently used methods in outcome and impact evaluation studies. Based on a combination of comparisons before and after the intervention as well as comparisons of the treatment and comparison groups, the method has an intuitive appeal and has been widely used in economics, public policy, health research, management, and other fields [37,38]. Thus, we employed the DID estimation technique to measure the effectiveness of the intervention using data from before and after the intervention in comparison of intervention and comparison groups. It was applied predominantly to quantify and test whether the level of change in the outcome of interest in the intervention group was significant compared to the comparison group.

The DID approach applied a linear regression model and calculated the change over time in intervention and comparison

groups. It double-differenced the change over time in the intervention group compared to the comparison group. The method also generated a valid estimate of the causal effect if the implementation of the CBMP was the only factor that might cause a change in the association between the CBMP and average information use before and after the intervention, as shown in the equation below:

$$Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \times P_t) + \beta_4 (T_g \times P_t \times \text{covariates}) + \varepsilon_t$$

where Y_{gt} is the average level of information use, β_1 is the average difference in Y between the two groups that is common in both time periods, β_2 is the average change in Y from the baseline to the endpoint time period that is common to both groups, β_3 is the average change in Y from the baseline to the endpoint time period of the intervention group compared to the comparison group, and β_4 is the triple difference adjusted for some covariates.

To estimate the average change in information use over time using the DID model, we created a dummy variable by assigning 1 to the intervention group and 0 to the comparison group. Moreover, the preintervention period and postintervention period were assigned 0 and 1, respectively. Subsequently, we employed a difference-in-difference-in-differences (DIDID) method to assess whether the program effects were heterogeneous across sex (male vs female), age (≤30 years vs >30 years), educational level (diploma vs above diploma), location (rural vs urban), salary (≤5000 ETB [Ethiopian birr] vs >5000 ETB; a currency exchange rate of 44.32 ETB=US $1 is applicable), and experience (≤5 years vs >5 years) [15].

The model provided information about the β coefficients, *P* values, and 95% CIs. If the coefficient for the interaction term (ie, the DID estimator) was significant at an α value of .05, we determined that the intervention was responsible for causing the change in the treatment group. In addition, the coefficients for the triple difference (ie, the DIDID estimator) were examined using the covariates listed above; one group was judged as having benefited more than the other if the model provided significant β coefficients [37,38].

Ethics Approval

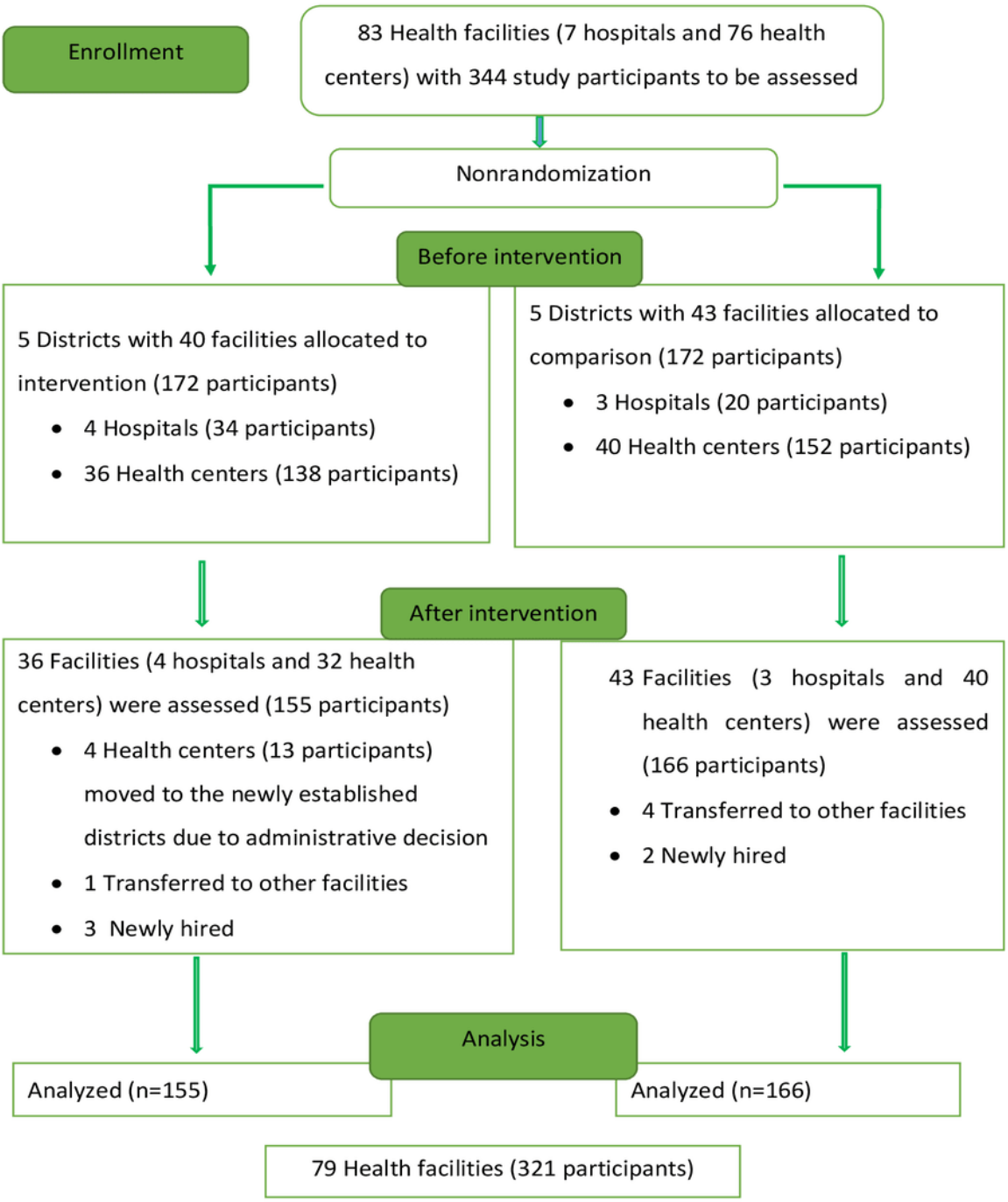
The study protocol was developed considering the ethical principles from the Declaration of Helsinki. The research protocol was registered at the Pan African Clinical Trials Registry (PACTR202001559723931), which is a World Health Organization International Clinical Trials Registry Platform primary register [39]. Moreover, the registry confirmed that the intervention was implemented with ethical consideration to human subject involvement. The CBMP offers original insights and is a new approach; the tailored intervention was implemented in an adaptive way to address the gaps identified at a specific intervention site. We also secured an ethical clearance letter from the UoG Institutional Review Board (reference No. O/V/P/RCS/05/430/2018). Participants were informed of the purpose of the study and consented before any inquiry. Data were collected anonymously with no personal identifiers; data were used only for this study. We presented findings with no manipulation or subject involvement.

Results

Participant Flow Through the Study

Baseline data were collected from 344 study participants (intervention: n=172, 50%; comparison: n=172, 50%) across 83 health facilities. However, a total of 321 study participants (intervention: n=155, 48.3%; comparison: n=166, 51.7%) across 79 health facilities (intervention: n=36, 46%; comparison: n=43, 54%) were surveyed at the endpoint of the study; the response rate was 93.3% (321/344). A total of 4 facilities out of 40 (10%) from the intervention arm were excluded because they became part of the newly established districts (Figure 2).

Figure 2. Flow diagram of study participants in the Amhara region, Northwest Ethiopia, 2021.



Characteristics of Study Participants

Among the 321 total study participants, 155 (48.3%) were from the intervention districts and 166 (51.7%) were from the comparison districts. More than half of the study participants were male in both intervention and comparison districts at baseline and endpoint periods. Almost two-thirds of the study participants were below the age of 28 years in both arms. Similarly, more than half of the study participants were diploma holders and resided in rural locations. Two-thirds of the participants earned equal to or below 5000 ETB at baseline, and nearly half of them earned above 5000 ETB at the study endpoint in both the intervention and comparison groups. More than half of the study participants had 5 years or more of experience (Table 1).

Table 1. Sociodemographic characteristics of study participants in the Amhara region, Northwest Ethiopia, 2021.

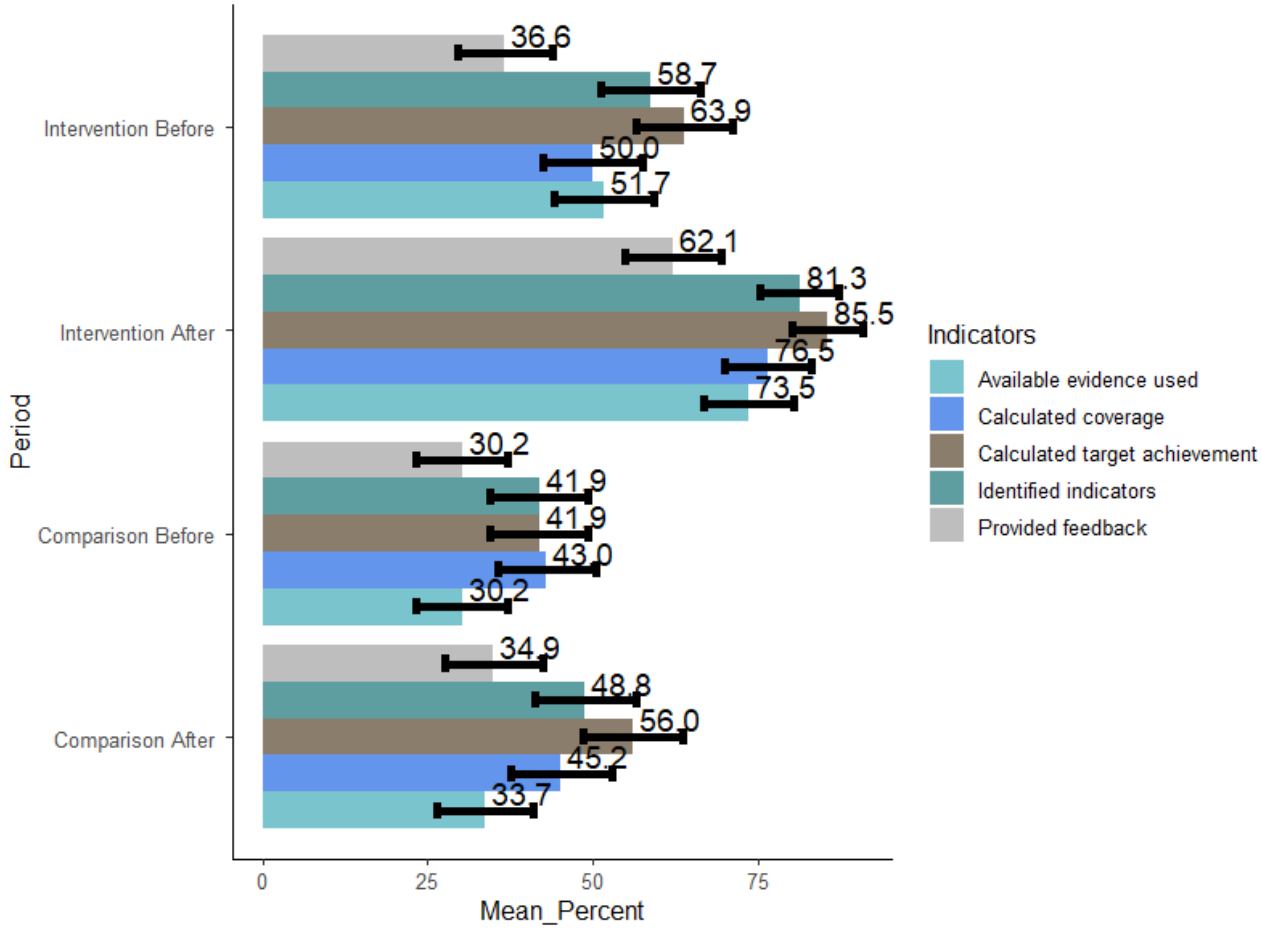
Variable	Baseline				Endpoint			
	Intervention		Comparison		Intervention		Comparison	
	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI	n (%)	95% CI
Sex								
Female	68 (39.5)	32.3-47.3	63 (36.6)	29.5-44.3	68 (43.9)	35.9-52.1	60 (36.4)	29.1-44.2
Male	104 (60.5)	52.7-67.7	109 (63.4)	55.7-70.5	87 (56.1)	47.9-64.0	106 (63.6)	55.8-70.9
Age (years)								
≤30	133 (77.3)	70.2-83.3	147 (85.5)	79.1-90.2	129 (83.2)	76.2-88.6	130 (78.3)	71.1-84.2
>30	39 (22.7)	16.8-29.8	25 (14.5)	9.8-20.9	26 (16.8)	11.4-23.8	36 (21.7)	15.8-28.9
Location								
Rural	94 (54.7)	46.9-62.2	122 (70.9)	63.4-77.5	80 (51.6)	43.5-59.7	116 (69.9)	62.2-76.6
Urban	78 (45.3)	37.8-53.1	50 (29.1)	22.5-36.6	75 (48.4)	40.3-56.5	50 (30.1)	23.4-37.8
Educational level								
Diploma or below	104 (60.5)	52.7-67.7	97 (56.4)	48.6-63.9	87 (56.1)	47.9-64.0	90 (54.5)	46.6-62.2
Above diploma	68 (39.5)	32.3-47.3	75 (43.6)	36.1-51.4	68 (43.9)	35.9-52.0	75 (45.5)	37.8-53.4
Salary (ETB[a])								
≤5000	131 (76.2)	68.9-82.2	109 (65.3)	57.5-72.4	77 (49.7)	41.6-54.6	80 (48.5)	40.7-56.4
>5000	41 (23.8)	17.8-31.0	58 (34.7)	27.6-42.5	78 (50.3)	42.2-58.4	85 (51.5)	43.6-59.3
Experience (years)								
≤5	107 (62.9)	55.2-70.1	128 (65.3)	58.1-71.9	89 (57.4)	49.2-65.2	97 (58.8)	50.9-66.3
>5	63 (37.1)	29.9-44.8	68 (34.7)	28.1-41.9	66 (42.6)	34.8-50.8	68 (41.2)	33.7-49.1

[a]ETB: Ethiopian birr; a currency exchange rate of 44.32 ETB=US $1 is applicable.

Component Indicators of Information Use

The study indicated that a mean of 30.2% (95% CI 23.3-37.2) and 51.7% (95% CI 44.23-59.3) of the department heads in comparison and intervention districts, respectively, used available evidence while making decisions at baseline, whereas a mean of 33.7% (95% CI 26.4-41.0) and 73.5% (95% CI 66.7-80.3) of the department heads in comparison and intervention districts, respectively, used available evidence while making decisions at study endpoint. At baseline, a mean of 41.9% (95% CI 34.5-49.3) and 43.0% (95% CI 35.5-50.5) of the departments in the comparison group calculated target achievement and program coverage, respectively, whereas a mean of 63.9% (95% CI 56.7-71.2) and 50.0% (95% CI 42.5-57.5) of the departments in the intervention group did so. However, the postperiod data showed that a mean of 56.0% (95% CI 48.4-63.7) and 45.2% (95% CI 37.6-52.8) of the comparison groups calculated target achievement and program coverage, respectively, whereas a mean of 85.5% (95% CI 80.1-90.9) and 76.5% (95% CI 70.0-83.0) of the intervention groups did so. At baseline, less than half of the study participants in both groups provided feedback to health workers at the lower levels; however, at the study endpoint, a mean of 34.9% (95% CI 27.5-42.3) of comparison group participants and 62.1% (95% CI 54.6-69.5) of the intervention group participants did so. At baseline, a mean of 41.9% (95% CI 34.5-49.3) and 58.7% (95% CI 51.3-66.2) of the departments in the comparison and intervention groups, respectively, had identified indicators, whereas at the study endpoint, a mean of 48.8% (95% CI 41.1-56.5) and 81.3% (95% CI 75.3-87.3) of the departments in the comparison and intervention groups, respectively, had done so (Figure 3).

Figure 3. Component indicators of routine information use at baseline and at the study endpoint in the comparison and intervention districts in the Amhara region, Northwest Ethiopia, 2021. The mean values are reported on the bars; the whiskers represent 95% CI values.



Average Level of Information Use in Decision-making

The average level of information use for the comparison group was 37.3% (95% CI 31.1%-43.6%) at baseline and 43.7% (95% CI 37.9%-49.5%) at the study endpoint. The average level of information use for the intervention group was 52.2% (95% CI 46.2%-58.3%) at baseline and 75.8% (95% CI 71.6%-80.0%) at the study endpoint. The DID analysis indicated that the net program effect change over time was significant (*P*=.003). It indicated that the intervention resulted in a 17% (95% CI 5%-28%) increment in the level of information use among the intervention districts compared to the comparison districts (Table 2).

Table 2. DID analysis in control and intervention districts in the Amhara region, Northwest Ethiopia, 2021.

Parameter	Program effect size	95% CI	*P* value
Intercept	0.356	0.300-0.413	<.001
Group (intervention vs comparison)	0.162	0.080-0.243	<.001
Time (study endpoint vs baseline)	0.08	0.000-0.160	.047
DID[a] analysis (group × time)	0.173	0.058-0.288	.003

[a]DID: difference-in-differences.

Subgroup Analysis for Program Effect Heterogeneity Diagnosis

The DIDID estimate showed that the CBMP increased information use by 16% among department heads who were working in urban facilities, with a DIDID estimate of 0.16 (95% CI 0.026-0.289; *P*=.02). However, we did not find much evidence of heterogeneity in program effect based on sex, age, educational level, salary, and experience, with DIDID estimates of 0.046 (95% CI –0.089 to 0.182), –0.002 (95% CI –0.015 to 0.009), –0.055 (95% CI –0.190 to 0.079), –1.63 (95% CI –5.22 to 1.95), and –0.006 (95% CI –0.017 to 0.005), respectively (Table 3).

Table 3. Subgroup analysis for selected variables in assessing program effect heterogeneity in the Amhara region, Northwest Ethiopia, 2021.

Effect modifier	Heterogeneity in program effect (95% CI)	*P* value
Sex (male vs female)	0.046 (–0.089 to 0.182)	.50
Age (≤30 years vs >30 years)	–0.002 (–0.015 to 0.009)	.65
Educational level (diploma vs above diploma)	–0.055 (–0.190 to 0.079)	.42
Salary (≤5000 ETB[a] vs >5000 ETB)	–0.00016 (–0.0005 to 0.00019)	.37
Residence (urban vs rural)	0.16 (0.026 to 0.289)	.02
Experience (≤5 years vs >5 years)	–0.006 (–0.017 to 0.005)	.29

[a]ETB: Ethiopian birr; a currency exchange rate of 44.32 ETB=US $1 is applicable.

Discussion

Principal Findings

All five component indicators showed high improvement at the study endpoint compared to baseline in each group. The intervention resulted in a 17% change in the average level of information use among study participants in the intervention districts compared to the comparison districts. We noted that the effect of the intervention was heterogeneous in urban and rural facilities. However, significant differences were not observed based on the sex, age, educational level, salary, and experience of the study participants.

This research revealed that the CBMP was effective in improving the capacity of the department and facility heads in using routine health information for decision-making and action. This finding was higher than that found in a cluster randomized controlled trial conducted in Sierra Leone on a community health data review meeting, which resulted in a 14% increment in evidence generation [40]. However, this finding was by far below the findings reported in a study conducted in Nigeria; in that study, data quality and information use training were found to improve feedback mechanisms by 54% [32]. This difference could be due to the large number of facilities covered by the CBMP and the nature of the study participants. Building health workers' capacity in data use for actions at all levels in the health system would improve and make more efficient the use of health care resources, which would, in turn, lead to making quality services available to clients.

As indicated with these findings, the intervention resulted in the majority of study participants having identified and used indicators to track performance progress in their catchment area. It was consistent with a single study done in Zanzibar and Tanzania where the data use workshop resulted in improvement in the local use of target indicators [10]. Identifying and using indicators in the health system enable health workers to measure the occurrence of disease or other health conditions and factors contributing to them [41]. In addition, indicators link information to actions and provide signals as to whether a program is effective and efficient in achieving the intended results in the target groups [42].

Though the findings highlighted an improvement in providing feedback to health workers at lower levels, it was still unsatisfactory compared to the desired level [27]; however, it was better than the findings of the study done in the Southern Nations, Nationalities, and Peoples Region [43]. It was also inconsistent with the findings obtained in Egypt that reported the effectiveness of the Feedback and Analytic Comparison Tool intervention in improving clinicians' capacity on providing feedback [12]. The difference might be because the latter was a single intervention primarily targeted at improving feedback mechanisms. In addition, it could be associated with low attention given to the importance of feedback among the study participants in our study. Generating synthesized evidence that indicates the strengths and weaknesses of health workers in the health system is one of the solemn expected activities, among others, in realizing the information revolution agenda [44]. Thus, ineffective feedback mechanisms lead to the provision of poor-quality services to clients and patients.

It was evidenced in a subgroup analysis that urban dwellers benefited more from the intervention than their counterparts. This finding is also in line with the baseline study finding, which indicated that work location was a significant factor associated with the level of information use [15]. This may be because more senior and qualified health workers had transferred from rural to urban facilities. Staff transfer is a common practice in the health system to reduce staff attrition rate [45]. This imbalance created different achievement levels in reaching the information revolution targets in all districts, which, in turn, can affect the quality of care provided to beneficiaries in general.

Limitations

One limitation of this study was that we did not employ randomization and blinding because of the nature of the study. This may have introduced some information leakage between intervention and comparison districts. Some facilities from which we took baseline data were not included in the endpoint survey, which may have biased the results. Moreover, social desirability and recall bias may also have been introduced, since participants were part of the intervention.

Conclusions

The CBMP was found to be effective in improving department and facility heads' capacity in using routine health information for decision-making. The intervention was more beneficial to study participants who resided in urban facilities than their counterparts. A remarkable change was observed in using the available evidence to inform decisions, identify indicators for tracking performance progress, compare targets versus achievements, and calculate program coverage. However, there is still a gap in providing synthesized feedback to health workers

at lower levels. Therefore, we propose the following recommendations. The intervention has been proven to produce positive outcomes and should be scaled up to other districts. Moreover, attention should be given to enhance the capacity of health workers employed in rural facilities and to strengthen feedback mechanisms at all levels, in order to reach the desired outcomes of the information revolution.

Acknowledgments

The research team members are thankful to the University of Gondar for providing an ethical review letter. We offer our appreciation to the Amhara Regional Health Bureau and the Amhara Public Health Institute for providing us with the support letter to conduct the research. The district office, health facilities, and departments were also supportive of our data collection, and we offer our gratitude. In the end, we are also grateful to the study participants and data collectors for their cooperation and involvement in the study.

This work would not be possible without the financial support from the Doris Duke Charitable Foundation (grant 2017187) through the University of Gondar. The funder had no role in data design, data collection, analysis, and interpretation of the study findings. The mission of the Doris Duke Charitable Foundation is to improve the quality of people's lives through grants supporting the performing arts, environmental conservation, medical research, and child well-being, and through preservation of the cultural and environmental legacy of Doris Duke's properties. We would also like to acknowledge the Norwegian Partnership Programme for Global Academic Cooperation for its financial contribution.

Data Availability

Data are available upon reasonable request from the corresponding author.

Authors' Contributions

MAC participated in data analysis. All authors conceptualized the design of the study, provided a review of the methodology and analysis results, contributed to the writing of the manuscript, and read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Almunawar MN, Anshari M. Health information systems (HIS): Concept and technology. ArXiv. Preprint posted online on March 18, 2012 [FREE Full text]
2. Hodge N. What are health information systems, and why are they important? Pac Health Dialog 2012 Apr;18(1):15-19. [Medline: 23240331]
3. Greenes RA. Health information systems 2025. In: Weaver CA, Ball MJ, Kim GR, Kiel JM, editors. Healthcare Information Management Systems: Cases, Strategies, and Solutions. 4th edition. Cham, Switzerland: Springer; 2016:579-600.
4. Fichman R, Kohli R, Krishnan R. The role of information systems in healthcare: Current research and future trends. Inf Syst Action Res 2011;22(3):28. [doi: 10.1007/978-0-387-36060-7_10]
5. Kolodner RM, Cohn SP, Friedman CP. Health information technology: Strategic initiatives, real progress. Health Aff (Millwood) 2008 Jan;27(5):w391-w395. [doi: 10.1377/hlthaff.27.5.w391] [Medline: 18713825]
6. Ismail S, Alshmari M, Latif K, Ahmad HF. A granular ontology model for maternal and child health information system. J Healthc Eng 2017;2017:9519321 [FREE Full text] [doi: 10.1155/2017/9519321] [Medline: 29065669]
7. Beane A, De Silva AP, Athapattu PL, Jayasinghe S, Abayadeera AU, Wijerathne M, et al. Addressing the information deficit in global health: Lessons from a digital acute care platform in Sri Lanka. BMJ Glob Health 2019;4(1):e001134 [FREE Full text] [doi: 10.1136/bmjgh-2018-001134] [Medline: 30775004]
8. Mutale W, Chintu N, Amoroso C, Awoonor-Williams K, Phillips J, Baynes C, et al. Improving health information systems for decision making across five sub-Saharan African countries: Implementation strategies from the African Health Initiative. BMC Health Serv Res 2013 May 31;13(S2):1-12. [doi: 10.1186/1472-6963-13-s2-s9]
9. Wagenaar BH, Hirschhorn LR, Henley C, Gremu A, Sindano N, Chilengi R, AHI PHIT Partnership Collaborative. Data-driven quality improvement in low-and middle-income country health systems: Lessons from seven years of implementation experience across Mozambique, Rwanda, and Zambia. BMC Health Serv Res 2017 Dec 21;17(Suppl 3):830 [FREE Full text] [doi: 10.1186/s12913-017-2661-x] [Medline: 29297319]
10. Braa J, Heywood A, Sahay S. Improving quality and use of data through data-use workshops: Zanzibar, United Republic of Tanzania. Bull World Health Organ 2012 May 01;90(5):379-384. [doi: 10.2471/blt.11.099580]
11. Mpofu M, Semo B, Grignon J, Lebelonyane R, Ludick S, Matshediso E, et al. BMC Public Health 2014 Oct 03;14:1032 [FREE Full text] [doi: 10.1186/1471-2458-14-1032] [Medline: 25281354]

12. Gaumer G, Hassan N, Murphy M. A simple primary care information system featuring feedback to clinicians. Int J Health Plann Manage 2008;23(3):185-202. [doi: 10.1002/hpm.899] [Medline: 17853507]
13. Nutley T, McNabb S, Salentine S. Impact of a decision-support tool on decision making at the district level in Kenya. Health Res Policy Syst 2013 Sep 08;11:34 [FREE Full text] [doi: 10.1186/1478-4505-11-34] [Medline: 24011028]
14. Bradley E, Hartwig KA, Rowe LA, Cherlin EJ, Pashman J, Wong R, et al. Hospital quality improvement in Ethiopia: A partnership-mentoring model. Int J Qual Health Care 2008 Dec;20(6):392-399. [doi: 10.1093/intqhc/mzn042] [Medline: 18784268]
15. Chanyalew M, Yitayal M, Atnafu A, Tilahun B. Routine health information system utilization for evidence-based decision making in Amhara national regional state, northwest Ethiopia: A multi-level analysis. BMC Med Inform Decis Mak 2021 Jan 26;21(1):28 [FREE Full text] [doi: 10.1186/s12911-021-01400-5] [Medline: 33499838]
16. Hoxha K, Hung YW, Irwin BR, Grépin KA. Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: A systematic review. Health Inf Manag 2020 Jun 30. [doi: 10.1177/1833358320928729] [Medline: 32602368]
17. Wude H, Woldie M, Melese D, Lolaso T, Balcha B. Utilization of routine health information and associated factors among health workers in Hadiya Zone, Southern Ethiopia. PLoS One 2020;15(5):e0233092 [FREE Full text] [doi: 10.1371/journal.pone.0233092] [Medline: 32437466]
18. Dagnew E, Woreta SA, Shiferaw AM. Routine health information utilization and associated factors among health care professionals working at public health institution in North Gondar, Northwest Ethiopia. BMC Health Serv Res 2018 Sep 04;18(1):685 [FREE Full text] [doi: 10.1186/s12913-018-3498-7] [Medline: 30180897]
19. Brattig Correia R, Chiodini J, Dalfovo O, Silva LH, Teske R. The use of information systems in health care facilities: A Brazilian case. J Technol Manag Innov 2013;8:143-144 [FREE Full text] [doi: 10.4067/s0718-27242013000300072]
20. Mate KS, Bennett B, Mphatswe W, Barker P, Rollins N. Challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa. PLoS One 2009;4(5):e5483 [FREE Full text] [doi: 10.1371/journal.pone.0005483] [Medline: 19434234]
21. Dondorp AM, Iyer SS, Schultz MJ. Critical care in resource-restricted settings. JAMA 2016 Feb 23;315(8):753-754. [doi: 10.1001/jama.2016.0976] [Medline: 26903331]
22. Rhatigan Jr JJ. Health systems and health care delivery. In: Ryan ET, Hill DR, Solomon T, Aronson NE, Endy TP, editors. Hunter's Tropical Medicine and Emerging Infectious Diseases. 10th edition. Amsterdam, the Netherlands: Elsevier; 2020:214-218.
23. Campbell DT, Stanley JC. Experimental and Quasi-Experimental Designs for Research. Boston, MA: Houghton Mifflin Company; 1963.
24. Central Statistics Agency. Population projection. Ethiopian Statistics Service. 2019. URL: https://www.statsethiopia.gov.et/population-projection/ [accessed 2022-04-11]
25. Ayalew D, Tesfaye K, Mamo G, Yitaferu B, Bayu W. Outlook of future climate in northwestern Ethiopia. Agric Sci 2012;3(4):608-624. [doi: 10.4236/as.2012.34074]
26. 2012 EFY Annual Report. Bahir-Dar, Ethiopia: Amhara Region Health Bureau; 2020.
27. Capacity Building and Mentorship Program (CBMP): Guiding Plan. Gondar, Ethiopia: University of Gondar; 2013.
28. Capacity Building and Mentorship Program (CBMP): Progress Report. Gondar, Ethiopia: University of Gondar; 2012.
29. Poulsen KM. Mentoring programmes: Learning opportunities for mentees, for mentors, for organisations and for society. Ind Commer Train 2013;45(5):63. [doi: 10.1108/ict-03-2013-0016]
30. Aqil A, Lippeveld T, Hozumi D. PRISM framework: A paradigm shift for designing, strengthening and evaluating routine health information systems. Health Policy Plan 2009 May;24(3):217-228 [FREE Full text] [doi: 10.1093/heapol/czp010] [Medline: 19304786]
31. Whitley E, Ball J. Statistics review 4: Sample size calculations. Crit Care 2002 Aug;6(4):335-341 [FREE Full text] [doi: 10.1186/cc1521] [Medline: 12225610]
32. Nwankwo B, Sambo M. Can training of health care workers improve data management practice in health management information systems: A case study of primary health care facilities in Kaduna State, Nigeria. Pan Afr Med J 2018;30:289 [FREE Full text] [doi: 10.11604/pamj.2018.30.289.15802] [Medline: 30637073]
33. Fox N, Amanda H, Mathers N. Sampling and Sample Size Calculation.: The NIHR RDS for the East Midlands / Yorkshire & the Humber; 2007. URL: https://www.bdct.nhs.uk/wp-content/uploads/2019/04/Sampling-and-Sample-Size-Calculation.pdf [accessed 2022-05-10]
34. Abera E, Daniel K, Letta T, Tsegaw D. Utilization of health management information system and associated factors in Hadiya zone health centers, Southern Ethiopia. Res Health Sci 2016 Sep 05;1(2):98. [doi: 10.22158/rhs.v1n2p98]
35. Aqil A, Hozumi D, Lippeveld T. PRISM Tools for Assessing, Monitoring, and Evaluating RHIS Performance. 2010 Mar. URL: https://www.measureevaluation.org/resources/publications/ms-09-34/at_download/document [accessed 2022-05-10]
36. Bonett DG, Wright TA. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. J Organ Behav 2014 Oct 13;36(1):3-15. [doi: 10.1002/job.1960]
37. Fredriksson A, de Oliveira GM. Impact evaluation using difference-in-differences. RAUSP Manage J 2019;54(4):519-532 [FREE Full text] [doi: 10.1108/RAUSP-05-2019-0112]

38. Lechner M. The estimation of causal effects by difference-in-difference methods. Found Trends Econom 2011;4(3):165-224. [doi: 10.1561/0800000014]
39. Chanyalew M. Effectiveness of health information system (HIS) training and mentorship on the level of information use among health facility and department heads: A quasi-experimental study. Pan African Clinical Trials Registry. 2020. URL: https://pactr.samrc.ac.za/TrialDisplay.aspx?TrialID=9602 [accessed 2022-05-10]
40. Cummings O'Connor E, Hutain J, Christensen M, Kamara M, Conteh A, Sarriot E, et al. Piloting a participatory, community-based health information system for strengthening community-based health services: Findings of a cluster-randomized controlled trial in the slums of Freetown, Sierra Leone. J Glob Health 2019 Jun;9(1):010418 [FREE Full text] [doi: 10.7189/jogh.09.010418] [Medline: 30842881]
41. Pan American Health Organization. Health indicators: Building blocks for health situation analysis. Epidemiol Bull 2001 Dec;22(4):1-5 [FREE Full text] [Medline: 12058680]
42. HMIS Reform Team. Health Management Information System (HMIS) / Monitoring and Evaluation (M&E). Addis Ababa, Ethiopia: Federal Ministry of Health; 2008 Jan. URL: https://www.cmpethiopia.org/content/download/478/2765/file/Health%20Managment%20Information%20System%20(HMIS).pdf [accessed 2021-07-29]
43. Belay H, Azim T, Kassahun H. Assessment of Health Management Information System (HMIS) Performance in SNNPR, Ethiopia. 2014 May. URL: https://pdf.usaid.gov/pdf_docs/pa00k27k.pdf [accessed 2022-05-10]
44. Information Revolution Roadmap. Addis Ababa, Ethiopia: Ethiopian Federal Ministry of Health; 2016 Apr. URL: http://repository.iifphc.org/bitstream/handle/123456789/316/Information%20Revolution%20Roadmap.pdf?sequence=1&isAllowed=y [accessed 2022-04-10]
45. Atnafu K, Tiruneh G, Ejigu T. Magnitude and associated factors of health professionals' attrition from public health sectors in Bahir Dar City, Ethiopia. Health 2013;5(11):1909-1916. [doi: 10.4236/health.2013.511258]

Abbreviations

ARHB: Amhara Regional Health Bureau
CBMP: Capacity Building and Mentorship Program
DHIS2: District Health Information Software 2
DID: difference-in-differences
DIDID: difference-in-difference-in-differences
ETB: Ethiopian birr
FMoH: Federal Ministry of Health
HIS: health information system
LMIC: low- and middle-income countries
OBAT: Organizational and Behavioral Assessment Tool
PI: principal investigator
RHIS: routine health information system
UoG: University of Gondar

Edited by C Lovis; submitted 18.05.21; peer-reviewed by J Colquitt, Jr, I Mircheva; comments to author 31.01.22; revised version received 13.02.22; accepted 25.02.22; published 22.04.22.

Please cite as:
Chanyalew MA, Yitayal M, Atnafu A, Mengiste SA, Tilahun B
The Effectiveness of the Capacity Building and Mentorship Program in Improving Evidence-Based Decision-making in the Amhara Region, Northwest Ethiopia: Difference-in-Differences Study
JMIR Med Inform 2022;10(4):e30518
URL: https://medinform.jmir.org/2022/4/e30518
doi:10.2196/30518
PMID:35451990

Original Paper

Cluster Analysis of Primary Care Physician Phenotypes for Electronic Health Record Use: Retrospective Cohort Study

Allan Fong[1], MS; Mark Iscoe[2], MD; Christine A Sinsky[3], MD; Adrian D Haimovich[2], MD, PhD; Brian Williams[4], MD; Ryan T O'Connell[4], MD; Richard Goldstein[4], MD, PhD; Edward Melnick[2], MHS, MD

[1]National Center for Human Factors in Healthcare, MedStar Health, Washington, DC, United States

[2]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States

[3]American Medical Association, Chicago, IL, United States

[4]Northeast Medical Group, Yale New Haven Health, Stratford, CT, United States

Corresponding Author:
Allan Fong, MS
National Center for Human Factors in Healthcare
MedStar Health
3007 Tilden St NW
Washington, DC, 20008
United States
Phone: 1 2022449807
Email: allan.fong@medstar.net

Abstract

Background: Electronic health records (EHRs) have become ubiquitous in US office-based physician practices. However, the different ways in which users engage with EHRs remain poorly characterized.

Objective: The aim of this study is to explore EHR use phenotypes among ambulatory care physicians.

Methods: In this retrospective cohort analysis, we applied affinity propagation, an unsupervised clustering machine learning technique, to identify EHR user types among primary care physicians.

Results: We identified 4 distinct phenotype clusters generalized across internal medicine, family medicine, and pediatrics specialties. Total EHR use varied for physicians in 2 clusters with above-average ratios of work outside of scheduled hours. This finding suggested that one cluster of physicians may have worked outside of scheduled hours out of necessity, whereas the other preferred ad hoc work hours. The two remaining clusters represented physicians with below-average EHR time and physicians who spend the largest proportion of their EHR time on documentation.

Conclusions: These findings demonstrate the utility of cluster analysis for exploring EHR use phenotypes and may offer opportunities for interventions to improve interface design to better support users' needs.

(JMIR Med Inform 2022;10(4):e34954) doi:10.2196/34954

KEYWORDS

electronic health record; phenotypes; cluster analysis; unsupervised machine learning; machine learning; EHR; primary care

Introduction

As of 2021, the vast majority of US office-based physicians used an electronic health record (EHR) [1]. The transition from paper to electronic records has many potential benefits but has also introduced new burdens. Furthermore, EHR use dominates clinical time [2] and is associated with burnout [3-5]. Despite the ubiquity of EHRs, patterns of clinician use are poorly characterized.

A 2019 survey study of clinicians reported widely divergent, subjective experiences with their EHR use and found that individual user differences accounted for over half of the variation in EHR use [6]. User-level variation can be due to disparities in proficiency that could potentially be remedied with appropriate training [7-10]. Emerging evidence suggests there are elements aside from proficiency that differentiate EHR users. For example, recent cross-sectional analyses of ambulatory care physicians' EHR use have found significant differences in time spent on EHRs based on gender [11,12], specialty [12,13], and country [14].

Audit logs offer a wealth of information derived from granular observations of users' EHR actions [15,16]. For example, research using log data has demonstrated associations between physicians' EHR activities and vendor-defined metrics of efficiency [17] and that efficiency varied based on physicians' years of experience and shift type [18]. In this study, we propose to use audit log data for the de novo identification of EHR user types (ie, EHR use phenotypes). Phenotype was first introduced by Richesson et al [19] as a biological concept to describe a set of observable biological traits. In the context of EHR use measures, phenotype will be used to describe observable use patterns across gender and specialty differences as defined by an unsupervised clustering approach called affinity propagation. First, 5 EHR use measures will be standardized using z-scores, which will then be used to calculate the similarities between physicians. A grid search and algorithm constraints will then be used to identify optimal clusters across a cohort of ambulatory care physicians.

Methods

Study Setting and Data Sources

This study retrospectively examined EHR log data of nontrainee, primary care physicians employed by a large ambulatory practice network (Northeast Medical Group) in northeastern United States (Connecticut, New York, and Rhode Island) between March 2018 and February 2020. Physicians were included if they specialized in general internal medicine, family medicine, or general pediatrics.

Ethics Approval

All data were anonymized, with the investigators blinded to the participants' identities. The study protocol was approved by Northeast Medical Group's Institutional Review Board (IRB number 2000026556).

EHR Use Measures

We retrieved data from the Epic Signal platform (Epic Systems) stratified by month and derived 5 proposed, time-based core EHR use measures normalized to 8 hours of scheduled patient time (Table 1) [20]. The first measure is EHR-Time$_8$, defined as the time a physician spends on EHRs (both during and outside of scheduled patient hours) [20]. The second measure is work outside of work (WOW$_8$), not to be confused with WOW carts (ie, workstations on wheels, a common industry term). WOW$_8$ is defined as the time a physician works on EHRs outside of scheduled patient hours [20]. The third measure is Note-Time$_8$, defined as the time a physician spends on documentation [20]. The fourth and fifth measures are IB-Time$_8$ and Order-Time$_8$, defined as the times a physician spends on inbox activities and on orders, respectively [20]. To account for relationships between EHR-Time$_8$ and its composite measures, we reported the ratios of WOW$_8$, Note-Time$_8$, IB-Time$_8$, and Order-Time$_8$ to EHR-Time$_8$, denoted as WOW-EHR, Note-EHR, IB-EHR, and Order-EHR, respectively. These measures (Table 1) were calculated and extracted from the Epic Signal platform, which have been validated and used in previous studies [20,21]. Each physician's EHR use measures were averaged across study months to account for variation in metric calculations introduced by changes in measure definitions over time due to the vendor's continuous quality improvement processes. For this analysis, we only considered physicians with valid metric months. Months with fewer than 30 clinical hours scheduled and less than 1 hour of EHR use were excluded from the analysis as invalid metric months. These thresholds were determined based on previous manual chart review validation and analysis of EHR vendor data [13].

Table 1. Electronic health record (EHR) use measures and definitions.

Measure	Definition
EHR-Time$_8$	Time a physician spends on EHRs (both during and outside of scheduled patient hours) normalized to 8 hours of scheduled patient time
WOW-EHR	Ratio of EHR time that occurs during work outside of work (WOW$_8$[a]) hours: WOW$_8$/EHR-Time$_8$
Note-EHR	Ratio of EHR time a physician spends on documentation: Note-Time$_8$[b]/EHR-Time$_8$
IB-EHR	Ratio of EHR time a physician spends on inbox (IB) activities: IB-Time$_8$[c]/EHR-Time$_8$
Order-EHR	Ratio of EHR time a physician spends on orders: Order-Time$_8$[d]/EHR-Time$_8$

[a]WOW$_8$: work outside of work hours normalized to 8 hours of scheduled patient time.

[b]Note-Time$_8$: note time hours normalized to 8 hours of scheduled patient time.

[c]IB-Time$_8$: inbox time hours normalized to 8 hours of scheduled patient time.

[d]Order-Time$_8$: order time hours normalized to 8 hours of scheduled patient time.

Cluster Analysis

Clusters were required to include individuals from at least two primary care specialties. Moreover, we did not require that all individuals be assigned to a phenotype cluster while also seeking to minimize the total number of phenotypes. Affinity propagation, an algorithm that takes a set of pairwise similarities between data points and finds clusters on the basis of maximizing the total similarity between data points in a cluster, was used for phenotype discovery [22]. Affinity propagation has advantages over other clustering algorithms, such as not predefining a number of clusters. A major disadvantage of

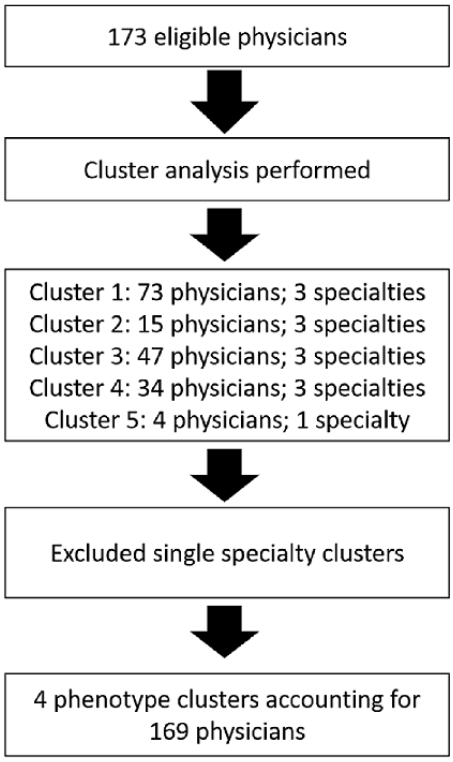
affinity propagation is its high computational cost and resource requirement; however, this approach was deemed feasible given this study's sample size [22]. First, a standard z-score for each measure was calculated in order to center and scale the data. Similarities between data points were then calculated using Euclidean distance, which is defined for two 2D points as the length of the line formed by the two points. A grid search was then performed by varying the damping factor and preference from 0.5 to 1 and from 2 to 4, respectively, to identify the optimal clustering given the initial cluster conditions. Physicians in clusters that did not have representation from at least two specialties were excluded. Finally, physician gender and specialty distributions were described between clusters. All analyses were performed using Python software (version 3.7; Python Software Foundation) and scikit-learn (version 0.24; scikit-learn developers) [23].

Results

Identifying Clusters

Of 332 ambulatory, nontrainee physicians, 290 (87.3%) have valid month metrics. Of those, a further 173 (52.1%) eligible physicians were of the specialties of interest: 117 (67.6%) in internal medicine, 36 (20.8%) in family medicine, and 20 (11.6%) in pediatrics. Gender distribution of the eligible physicians was 47.4% (82/173) female and 52.6% (91/173) male. We identified 4 clusters that met our a priori defined clustering conditions, accounting for 97.7% (169/173) of eligible physicians (Figure 1).

Figure 1. Summary of workflow and exclusion criteria.



EHR Use Measures and Phenotypes Clusters

The phenotype clusters are "Lower EHR time," "Higher note time," "Work outside of work," and "Notes outside of work." The EHR use measures across clusters are summarized in Table 2. There was a significant association between phenotype clusters and each EHR use measure: EHR-Time$_8$ (Kruskal-Wallis H=72.7, P<.001), WOW-EHR (H=84.3, P<.001), Note-EHR (H=89.0, P<.001), IB-EHR (H=45.8, P<.001), and Order-EHR (H=46.8, P<.001). The z-scores for the measures are displayed in Figure 2 to illustrate the relative differences between clusters.

Table 2. Electronic health record (EHR) use measures by phenotype cluster.

Measure	Phenotype clusters, median (IQR)				
	Lower EHR time	Higher note time	Work outside of work	Notes outside of work	All
EHR-Time$_8$[a]	4.62 (4.20-5.43)	5.81 (4.41-6.22)	6.83 (5.95-8.36)	5.90 (5.37-6.36)	5.62 (4.57-6.40)
WOW-EHR[b]	0.07 (0.04-0.12)	0.05 (0.03-0.07)	0.21 (0.17-0.26)	0.13 (0.10-0.19)	0.11 (0.06-0.19)
Note-EHR[c]	0.24 (0.20-0.28)	0.46 (0.43-0.49)	0.31 (0.27-0.36)	0.37 (0.33-0.40)	0.29 (0.24-0.38)
IB-EHR[d]	0.14 (0.12-0.18)	0.06 (0.05-0.08)	0.15 (0.11-0.17)	0.10 (0.08-0.12)	0.13 (0.09-0.16)
Order-EHR[e]	0.19 (0.17-0.24)	0.14 (0.12-0.17)	0.16 (0.14-0.18)	0.14 (0.12-0.17)	0.17 (0.14-0.20)

[a]EHR-Time$_8$: time a physician spends on EHRs normalized to 8 hours of scheduled patient time.

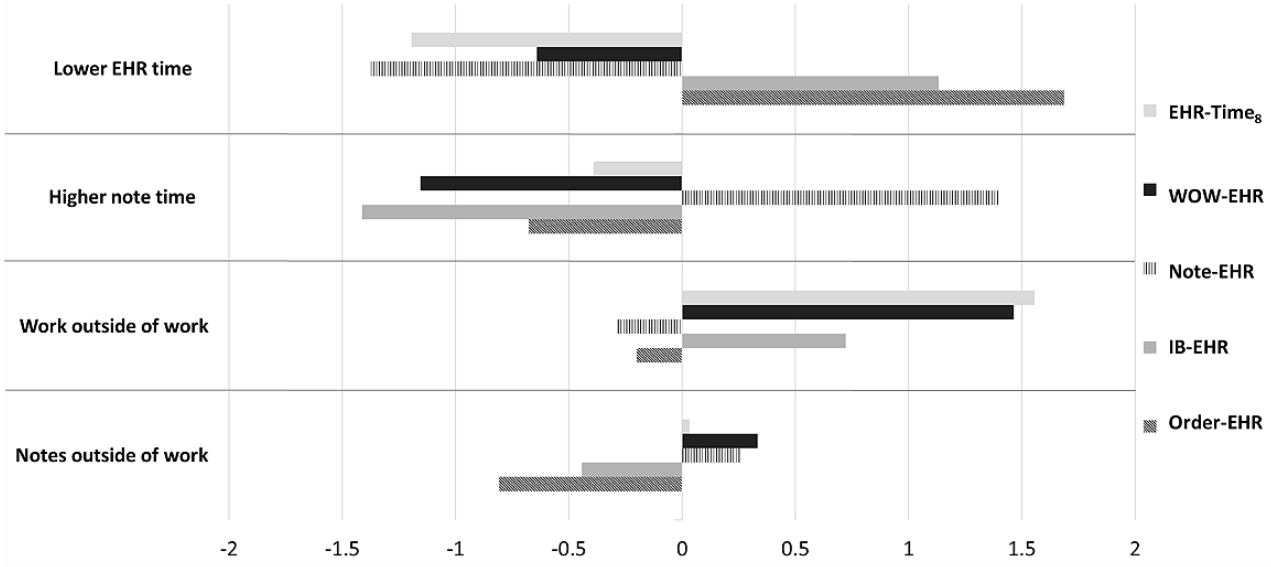
[b]WOW-EHR: ratio of EHR time that occurs during work outside of scheduled hours.

[c]Note-EHR: ratio of EHR time that a physician spends on documentation.

[d]IB-EHR: ratio of EHR time that a physician spends on inbox activities.

[e]Order-EHR: ratio of EHR time that a physician spends on orders.

Figure 2. Z-scores for electronic health record (EHR) use measure across clusters. EHR-Time$_8$: time a physician spends on EHRs normalized to 8 hours of scheduled patient time; IB-EHR: ratio of EHR time that a physician spends on inbox activities; Note-EHR: ratio of EHR time that a physician spends on documentation; Order-EHR: ratio of EHR time that a physician spends on orders; WOW-EHR: ratio of EHR time that occurs during work outside of scheduled hours.



"Lower EHR Time" Cluster

The "Lower EHR time" cluster was the largest cluster, constituting 42.2% (73/173) of eligible physicians. Physicians in this cluster spent the least amount of time on EHRs (EHR-Time$_8$: median 4.62, IQR 4.20-5.43). "Lower EHR time" cluster physicians had the lowest median Note-EHR ratio of 0.24 (IQR 0.20-0.28) and the second lowest median WOW-EHR ratio of 0.07 (IQR 0.04-0.12). They also had the highest median IB-EHR and Order-EHR ratios of 0.14 (IQR 0.12-0.18) and 0.19 (IQR 0.17-0.24), respectively.

"Higher Note Time" Cluster

"Higher note time" cluster physicians, constituting only 8.7% (15/173) of the total, had near-average normalized EHR time (EHR-Time$_8$: median 5.81, IQR 4.41-6.22). Physicians in this cluster spent the largest proportion of their EHR time documenting notes (Note-Time: median 0.46, IQR 0.43-0.49) compared to physicians in other clusters. They also spent the lowest proportions of that time on EHRs outside of scheduled hours and on inbox activities, with median WOW-EHR and IB-EHR ratios of 0.05 (IQR 0.03-0.07) and 0.06 (IQR 0.05-0.08), respectively.

"Work Outside of Work" Cluster

"Work outside of work" cluster physicians, constituting 27.2% (47/173) of the total, spent the most time on EHRs (EHR-Time$_8$: median 6.83, 5.95-8.36) and the largest proportion of that time outside of work hours (WOW-EHR: median 0.21, IQR 0.17-0.26). This cluster of physicians had average median Note-EHR and Order-EHR ratios of 0.31 (0.27-0.36) and 0.16 (IQR 0.14-0.18), respectively, and an above-average median IB-EHR ratio of 0.15 (IQR 0.11-0.17).

"Notes Outside of Work" Cluster

"Notes outside of work" cluster physicians, constituting 19.7% (34/173) of the total, had the second-highest median WOW-EHR ratio of 0.13 (IQR 0.10-0.19) but had near-average total normalized EHR time (EHR-Time$_8$: median 5.90, IQR 5.37-6.36). This cluster of physicians had an above-average median Note-EHR ratio of 0.37 (IQR 0.33-0.40) and below-average median IB-EHR and Order-EHR ratios of 0.10 (IQR 0.08-0.12) and 0.14 (IQR 0.12-0.17), respectively.

Phenotype Clusters by Specialty and Gender

Physician distribution across phenotype clusters by specialty and gender are reported in Table 3. There was a significant association between the clusters and specialty (X^2_6=26.67, P<.001). Pediatricians primarily fell into the "Higher note time" and "Notes outside of work" clusters (16/20, 80%) and accounted for 47% (7/15) of the total physicians in the "Higher note time" cluster. Family and internal medicine physicians were primarily distributed across the "Lower EHR time" and "Work outside of work" clusters (family medicine: 29/36, 81%; internal medicine: 87/113, 77%). In addition, there was a significant association between gender and clusters (X^2_3=18.28, P<.001). Female physicians were more prominent in the "Work outside of work" and "Notes outside of work" clusters, accounting for 64% (30/47) and 62% (21/34) of the clusters, respectively. Male physicians accounted for 71% (52/73) of the "Lower EHR time" cluster.

Table 3. Physician specialty and gender distribution by phenotype cluster.

Distribution	Number of physicians (N=173), n (%)	Phenotype clusters				*P* value
		Lower EHR[a] time (n=73), n (%)	Higher note time (n=15), n (%)	Work outside of work (n=47), n (%)	Notes outside of work (n=34), n (%)	
Specialty						<.001
Family medicine	36 (21)	19 (26)	2 (13)	10 (21)	5 (15)	
Internal medicine	113 (65)	52 (71)	6 (40)	35 (74)	20 (59)	
Pediatrics	20 (12)	2 (3)	7 (47)	2 (4)	9 (26)	
Gender						<.001
Female	80 (46)	21 (29)	8 (53)	30 (64)	21 (62)	
Male	89 (51)	52 (71)	7 (47)	17 (36)	13 (38)	
Total	169 (98)	73 (42)	15 (9)	47 (27)	34 (20)	

[a]EHR: electronic health record.

Discussion

Principal Findings

In this unsupervised clustering machine learning analysis of a cohort of primary care physicians, we identified 4 distinct EHR use phenotypes characterized by the total time spent on EHR activities and the ratios of those times in comparison to one another. These phenotypes were differentiated and described by patterns of use consistent with overall efficiency, higher documentation time, and working outside of work hours; each of these patterns of use were generally associated with the "Lower EHR time," "Higher note time," and "Work/Notes outside of work" clusters, respectively. While exploratory, these results provide insights into EHR use phenotypes across gender and specialties that can complement and provide additional context for current EHR use research.

Work Outside of Scheduled Hours

We identified 2 phenotype clusters that had above-average ratios for work outside of scheduled hours. Although "Work outside of work" and "Notes outside of work" clusters both had high WOW-EHR ratios, only the "Work outside of work" cluster had significantly higher than average EHR-Time$_8$. A possible explanation for this is that physicians in the "Work outside of work" cluster work from home partly out of necessity because they require more time on EHRs, whereas physicians in the "Notes outside of work" cluster may elect to finish work at home, suggesting a preference for ad hoc work hours.

Note Time

Time spent on clinical documentation accounted for the largest proportion of total EHR time in each cluster. There was, however, considerable variation in the ratio of note time to EHR time across clusters: from 0.24 of EHR time in the "Lower EHR time" cluster to 0.46 in the "Higher note time" cluster despite similar total EHR time in both clusters. Potential explanations for this variation include differences in clinic- or physician-specific workflows (eg, scribe support or team-based documentation; differences in depth and complexity of encounters and expectations for documentation; and use of form, copied, or auto-populated notes) and differences in documentation style, particularly among the "Higher note time" cluster that may include physicians who deliberately spend more time on documentation.

Limitations

This exploratory work only used time-based metrics and did not account for patient acuity or complexity. Although the data were gathered over a 2-year period, systemic differences in patient volume and care could have affected the results. In addition, this work was limited to a single ambulatory practice network in one region of the United States and was limited to primary care physicians. Some types of EHR activities (eg, chart review) were not included in the metrics, and it is possible that other activities or practice domains could also affect clustering. Furthermore, it should be noted that this study only identified EHR use phenotypes and did not explore reasons behind differences in EHR use or assign value to the phenotypes.

Conclusions

Our findings may highlight opportunities for interventions to improve EHR design and use to better support EHR users' needs. Potential differences in users' needs were identified for each phenotype cluster. The "Higher note time" and "Notes outside of work" clusters might benefit from scribe support more than the other two clusters. The "Work outside of work" cluster might benefit from inbox support and restructuring their practice for a more team-based approach. Physicians in the "Lower EHR time" cluster could be consulted as local champions to help their peers improve their EHR efficiency. By identifying and classifying individual EHR use and user needs, we can better understand and target interventions at the individual or department level. Future work should validate these phenotypes in larger cohorts and in diverse settings, explore differences in physicians' training and demographics across phenotypes, and investigate the relationships among EHR use phenotypes, patient outcomes, and clinician satisfaction and burnout.

Acknowledgments

We thank the physicians of the Northeast Medical Group. We also thank the following individuals for their contribution and support: Prem Thomas, MD, Center for Medical Informatics, Yale School of Medicine (data analysis, not compensated); and Becky Tylutki, Team Coordinator, Joint Data Analytics Team, Yale New Haven Health System (data collection and honest broker, not compensated). Financial support for this study was provided in part by two American Medical Association Practice Transformation Initiatives (contract numbers 36648 and 36650). The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report.

Conflicts of Interest

CAS is employed by the American Medical Association. All other authors declare no other conflicts of interest. The opinions expressed in this article are those of the authors and should not be interpreted as American Medical Association policy.

References

1. Office-based physician health IT adoption. The Office of the National Coordinator for Health Information Technology. 2021. URL: https://www.healthit.gov/data/apps/office-based-physician-health-it-adoption [accessed 2021-09-24]
2. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016 Dec 06;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
3. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
4. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
5. Olson K, Sinsky C, Rinne ST, Long T, Vender R, Mukherjee S, et al. Cross-sectional survey of workplace stressors associated with physician burnout measured by the Mini-Z and the Maslach Burnout Inventory. Stress Health 2019 Apr;35(2):157-175. [doi: 10.1002/smi.2849] [Medline: 30467949]
6. Longhurst CA, Davis T, Maneker A, Eschenroeder HC, Dunscombe R, Reynolds G, Arch Collaborative. Local investment in training drives electronic health record user satisfaction. Appl Clin Inform 2019 Mar;10(2):331-335 [FREE Full text] [doi: 10.1055/s-0039-1688753] [Medline: 31091545]
7. Robinson KE, Kersey JA. Novel electronic health record (EHR) education intervention in large healthcare organization improves quality, efficiency, time, and impact on burnout. Medicine (Baltimore) 2018 Sep;97(38):e12319 [FREE Full text] [doi: 10.1097/MD.0000000000012319] [Medline: 30235684]
8. Sieja A, Markley K, Pell J, Gonzalez C, Redig B, Kneeland P, et al. Optimization sprints: improving clinician satisfaction and teamwork by rapidly reducing electronic health record burden. Mayo Clin Proc 2019 May;94(5):793-802 [FREE Full text] [doi: 10.1016/j.mayocp.2018.08.036] [Medline: 30824281]

9. DiAngi YT, Stevens LA, Halpern-Felsher B, Pageler NM, Lee TC. Electronic health record (EHR) training program identifies a new tool to quantify the EHR time burden and improves providers' perceived control over their workload in the EHR. JAMIA Open 2019 Jul;2(2):222-230 [FREE Full text] [doi: 10.1093/jamiaopen/ooz003] [Medline: 31984357]
10. Hollister-Meadows L, Richesson RL, De Gagne J, Rawlins N. Association between evidence-based training and clinician proficiency in electronic health record use. J Am Med Inform Assoc 2021 Mar 18;28(4):824-831 [FREE Full text] [doi: 10.1093/jamia/ocaa333] [Medline: 33575787]
11. Tait SD, Oshima SM, Ren Y, Fenn AE, Boazak M, Hinz EM, et al. Electronic health record use by sex among physicians in an academic health care system. JAMA Intern Med 2021 Feb 01;181(2):288-290 [FREE Full text] [doi: 10.1001/jamainternmed.2020.5036] [Medline: 33284311]
12. Rotenstein LS, Holmgren AJ, Downing NL, Bates DW. Differences in total and after-hours electronic health record time across ambulatory specialties. JAMA Intern Med 2021 Jun 01;181(6):863-865 [FREE Full text] [doi: 10.1001/jamainternmed.2021.0256] [Medline: 33749732]
13. Melnick ER, Ong SY, Fong AL, Socrates V, Ratwani RM, Nath B, et al. Characterizing physician EHR use with vendor derived data: a feasibility study and cross-sectional analysis. J Am Med Inform Assoc 2021 Jul 14;28(7):1383-1392 [FREE Full text] [doi: 10.1093/jamia/ocab011] [Medline: 33822970]
14. Holmgren AJ, Downing NL, Bates DW, Shanafelt TD, Milstein A, Sharp CD, et al. Assessment of electronic health record use between US and non-US health systems. JAMA Intern Med 2021 Feb 01;181(2):251-259 [FREE Full text] [doi: 10.1001/jamainternmed.2020.7071] [Medline: 33315048]
15. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: A new goldmine for health services research? J Biomed Inform 2020 Jan;101:103343 [FREE Full text] [doi: 10.1016/j.jbi.2019.103343] [Medline: 31821887]
16. Amroze A, Field TS, Fouayzi H, Sundaresan D, Burns L, Garber L, et al. Use of electronic health record access and audit logs to identify physician actions following noninterruptive alert opening: descriptive study. JMIR Med Inform 2019 Feb 07;7(1):e12650 [FREE Full text] [doi: 10.2196/12650] [Medline: 30730293]
17. Kannampallil TG, Denton CA, Shapiro JS, Patel VL. Efficiency of emergency physicians: insights from an observational study using EHR log files. Appl Clin Inform 2018 Jan;9(1):99-104 [FREE Full text] [doi: 10.1055/s-0037-1621705] [Medline: 30184241]
18. Wang JK, Ouyang D, Hom J, Chi J, Chen JH. Characterizing electronic health record usage patterns of inpatient medicine residents using event log data. PLoS One 2019 Feb 06;14(2):e0205379 [FREE Full text] [doi: 10.1371/journal.pone.0205379] [Medline: 30726208]
19. 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 01;20(e2):e226-e231 [FREE Full text] [doi: 10.1136/amiajnl-2013-001926] [Medline: 23956018]
20. Sinsky CA, Rule A, Cohen G, Arndt BG, Shanafelt TD, Sharp CD, et al. Metrics for assessing physician activity using electronic health record log data. J Am Med Inform Assoc 2020 Apr 01;27(4):639-643 [FREE Full text] [doi: 10.1093/jamia/ocz223] [Medline: 32027360]
21. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. Ann Fam Med 2017 Sep;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
22. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007 Feb 16;315(5814):972-976. [doi: 10.1126/science.1136800] [Medline: 17218491]
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830. [doi: 10.48550/arXiv.1201.0490]

Abbreviations

EHR: electronic health record
EHR-Time$_8$: time a physician spends on EHRs normalized to 8 hours of scheduled patient time
IB-EHR: percent of EHR time that a physician spends on inbox activities
IB-Time$_8$: inbox time hours normalized to 8 hours of scheduled patient time
Note-EHR: percent of EHR time that a physician spends on documentation
Note-Time$_8$: note time hours normalized to 8 hours of scheduled patient time
Order-EHR: percent of EHR time that a physician spends on orders
Order-Time$_8$: order time hours normalized to 8 hours of scheduled patient time
WOW$_8$: work outside of work hours normalized to 8 hours of scheduled patient time
WOW-EHR: percent of EHR time that occurs during work outside of scheduled hours

Edited by C Lovis; submitted 14.11.21; peer-reviewed by R Verheij, A Joseph, M Pradhan; comments to author 08.01.22; revised version received 03.02.22; accepted 11.03.22; published 15.04.22.

Please cite as:
Fong A, Iscoe M, Sinsky CA, Haimovich AD, Williams B, O'Connell RT, Goldstein R, Melnick E
Cluster Analysis of Primary Care Physician Phenotypes for Electronic Health Record Use: Retrospective Cohort Study
JMIR Med Inform 2022;10(4):e34954
URL: https://medinform.jmir.org/2022/4/e34954
doi:10.2196/34954
PMID:35275070

Original Paper

The Factors Associated With Nonuse of and Dissatisfaction With the National Patient Portal in Finland in the Era of COVID-19: Population-Based Cross-sectional Survey

Emma Kainiemi[1], RM, MNSc; Tuulikki Vehko[1], PhD; Maiju Kyytsönen[1], RN, MHSc; Iiris Hörhammer[2], MSc, DSC; Sari Kujala[3], PhD; Vesa Jormanainen[1], MSc, MD; Tarja Heponiemi[1], PhD

[1]Finnish Institute for Health and Welfare, Helsinki, Finland

[2]Department of Industrial Engineering and Management, Aalto University, Espoo, Finland

[3]Department of Computer Science, Aalto University, Espoo, Finland

Corresponding Author:
Emma Kainiemi, RM, MNSc
Finnish Institute for Health and Welfare
PO Box 30
Mannerheimintie 166
Helsinki, 00271
Finland
Phone: 358 295248292
Email: emma.kainiemi@thl.fi

Abstract

Background: In the abnormal circumstances caused by the COVID-19 pandemic, patient portals have supported patient empowerment and engagement by providing patients with access to their health care documents and medical information. However, the potential benefits of patient portals cannot be utilized unless the patients accept and use the services. Disparities in the use of patient portals may exacerbate the already existing inequalities in health care access and health outcomes, possibly increasing the digital inequality in societies.

Objective: The aim of this study is to examine the factors associated with nonuse of and dissatisfaction with the Finnish nationwide patient portal My Kanta Pages among the users of health care services during the COVID-19 outbreak. Several factors related to sociodemographic characteristics, health, and the use of health care services; experiences of guidance concerning electronic services; and digital skills and attitudes were evaluated.

Methods: A national population survey was sent using stratified sampling to 13,200 Finnish residents who had reached the age of 20 years. Data were collected from September 2020 to February 2021 during the COVID-19 pandemic. Respondents who had used health care services and the internet for transactions or for searching for information in the past 12 months were included in the analyses. Bivariate logistic regression analyses were used to examine the adjusted associations of respondent characteristics with the nonuse of My Kanta Pages and dissatisfaction with the service. The inverse probability weighting (IPW) method was applied in all statistical analyses to correct for bias.

Results: In total, 3919 (64.9%) of 6034 respondents were included in the study. Most respondents (3330/3919, 85.0%) used My Kanta Pages, and 2841 (85.3%) of them were satisfied. Nonusers (589/3919, 15%) were a minority among all respondents, and only 489 (14.7%) of the 3330 users were dissatisfied with the service. Especially patients without a long-term illness (odds ratio [OR] 2.14, 95% CI 1.48-3.10), those who were not referred to electronic health care services by a professional (OR 2.51, 95% CI 1.70-3.71), and those in need of guidance using online social and health care services (OR 2.26, 95% CI 1.41-3.65) were more likely nonusers of the patient portal. Perceptions of poor health (OR 2.10, 95% CI 1.51-2.93) and security concerns (OR 1.87, 95% CI 1.33-2.62) were associated with dissatisfaction with the service.

Conclusions: Patients without long-term illnesses, those not referred to electronic health care services, and those in need of guidance on the use of online social and health care services seemed to be more likely nonusers of the Finnish nationwide patient portal. Moreover, poor health and security concerns appeared to be associated with dissatisfaction with the service. Interventions to promote referral to electronic health care services by professionals are needed. Attention should be targeted to information security of the service and promotion of the public's confidence in the protection of their confidential data.

(JMIR Med Inform 2022;10(4):e37500) doi:10.2196/37500

KEYWORDS

patient portal; digital technology; eHealth; nonuse; dissatisfaction; COVID-19; digital health; patient empowerment; epidemiology; population survey; national survey

Introduction

The worldwide COVID-19 pandemic limited the provision of nonurgent health care services [1,2]. During this time, the use and interest in patient portals increased [3,4] because portals have enabled patients to have continuous [5] and secure access to their health care documents and medical information [4]. Patient portals are electronic services that allow patients to access [6] and in some cases manage their electronic health record documentations [7] and interact with health care professionals [6,8,9]. The functionalities provided in a patient portal vary by portal and country [9,10].

Patient portals offer transparent information about the patients' health and well-being [11] and enhance the delivery of individualized care [3]. Patient portals have been reported to increase the patients' knowledge and understanding of their own health condition, and thus they might be better prepared for future contacts with health care professionals [12]. This supports patient empowerment and engagement [4,5] as the patients feel more involved and responsible for their own care [12]. In addition, patient portals might increase the patients' satisfaction with care [13-18] and improve patient safety [15,17]. However, high-quality evidence of health benefits has not yet been demonstrated [8].

The potential benefits of patient portals cannot be utilized unless the patients accept and adopt the service [7,19]. According to the information system (IS) success model, the benefits of using the service arise from its use and user satisfaction [20]. Further, increased user satisfaction will lead to increased use [20], and unmet expectations will alter the use and satisfaction with patient portals [21]. Not all the barriers related to the use of patient portals are related to practical issues, such as a lack of hardware and access to the internet, but patients may have other valid reasons for nonuse as well [7,22,23]. Patients are also in an unequal position in terms of using electronic health care services, since not everyone has the resources or the same possibilities to use the services and take more responsibility for the management of their own health and well-being [24].

Previous research has examined differences in patient portal use in different contexts and patient populations. Several studies have reported an association between portal use and sociodemographic background [6,7,14,23,25,26] and various health-related factors [7,12,14,25,27,28]. In addition, patients' guidance through increasing awareness and knowledge of patient portals [6,25,28], as well as endorsement and engagement with portals by health care professionals [7,25], have been identified as important associated factors. There are also some studies that have reported an association between the use of patient portals and factors related to the use of the internet, such as the frequency of use [26,29,30] and perceptions of the users' own internet skills [26,30]. Furthermore, it has been reported that perceptions of electronic services may encourage or impede their use [31]. However, these previous studies have been conducted under normal circumstances before the COVID-19 pandemic and are thus only partially applicable in the context changed by the pandemic [4].

Factors associated with the patients' satisfaction with patient portals have been less studied. Mainly descriptive research on the portal users' experiences exists, and only little research has been conducted with quantitative methods about factors associated with satisfaction [5]. Kong et al [5] examined factors that predicted portal use and the users' willingness to recommend the service among chronically ill patients during the COVID-19 pandemic in the Netherlands. They discovered that the respondent's level of control, hospital visit time, life satisfaction, and level of depression are significantly associated with portal use. Variables related to the portal user's waiting times for responses via the portal were the strongest predictors of the willingness to recommend the portal. However, the used variables concerned patients with long-term illnesses, and no comparisons were made to assess whether the same variables were associated with the use and willingness to recommend the patient portal. In addition, only little research exists on the factors associated with use and satisfaction using a nationally representative sample.

The aim of this study is to examine factors associated with the nonuse of and dissatisfaction with the Finnish nationwide patient portal My Kanta Pages (My Kanta) during the COVID-19 pandemic. Only respondents who had used the internet in the past 12 months were included to examine nonuse beyond the first-level digital divide [32] caused by a lack of hardware and access to the internet. Several factors related to (1) sociodemographic characteristics, (2) health and the use of health care services, (3) experiences of guidance concerning electronic services, and (4) digital skills and attitudes were examined. The evaluation of factors associated with the nonuse of and dissatisfaction with the patient portal is important to further develop the service and advocate for nonusers. Disparities in the use of patient portals might exacerbate the already existing disparities in health care access and health outcomes [33], increasing digital inequality in societies [34]. Knowledge of the factors that are associated with the nonuse of and dissatisfaction with the national patient portal in Finland, one of the pioneer countries in digitalization, can provide valuable information for countries and organizations that are further developing their electronic services.

Methods

Study Context

Finland is a sparsely populated country with 5.5 million residents. The health care system is decentralized, and until the end of 2022, municipalities (n=311) are responsible for

organizing health care services, which are funded by taxes, state transfers, and user fees [1]. Finland can be considered 1 of the leading countries in terms of digitalization [35]. One of the most widely used electronic service is the nationwide patient portal called My Kanta. Between the years 2010 and 2018, cumulatively 63% of the adult Finnish residents had accessed the service. There is a professional user interface for Kanta Services, which can be used by the public and private actors of the social welfare and health care sector [36]. In addition to the nationwide patient portal, some public and private actors offer their clients access to their own patient portals or regional portals [37].

My Kanta was launched step-by-step starting from 2010 [38] to promote patient safety as well as the continuity and transparency of care [39]. My Kanta enables continuous access of Finnish residents to their health information [9], including browsing their own electronic prescriptions and medical records, such as patient reports, laboratory results, and X-ray statements [40]. All producers of health care services have been obligated to use electronic prescriptions since 2017 [41], and patients can request prescription renewals via My Kanta [42]. In addition, it is possible to record and monitor well-being data, such as blood glucose or activity meter. To access My Kanta, a Finnish personal identity code is needed, and e-authorization must be made with identification using online banking codes, mobile identification, or a certificate card [42].

The number of My Kanta users has grown steadily since its launch [38], and the COVID-19 pandemic increased the number of logins because of the availability of coronavirus test results [43]. In 2020, the service was used 29.4 million times by a total of 2.7 million individual visitors [43]. E-prescriptions were issued approximately 26.4 million times during 2020 [44]. In the future, the use is expected to further increase as the deployment of authorization for an adult to act on behalf of another adult was introduced after the data collection period and authorization for guardians to act, with some restrictions, on behalf of their children aged 10-17 years [45] will be fully implemented.

Sample

This study was conducted in Finland as part of the FinSote 2020 National Survey of Health, Wellbeing, and Service Use [46]. The questionnaire was sent using stratified sampling to 13,200 Finnish residents who had reached the age of 20 years. Data were collected from September 2020 to February 2021 during the second wave of the COVID-19 pandemic. A possibility to respond either in electronic or in paper form in Finnish, Swedish, Russian, or English was offered. During the data collection, participants who had not responded were approached by mail up to 4 times.

Altogether, 6034 Finnish residents (n=3401 [56.4%] female, mean age 64.5 years, SD 17.9) responded to the questionnaire (response rate 46.5%). In total, 3919 (65.0%) respondents were included in the study sample as they had used health care services and the internet in the past 12 months. The sample was weighted using inverse probability weighting (IPW) correction [47]. The weights were estimated using sociodemographic register–based variables: the respondents' age, gender, marital status, level of education, area of residence, and native language. Information about the respondents' age, gender, and area of residence were obtained from the National Population Register. In previous research, the IPW method improved the accuracy of the results of a population survey and removed most of the bias caused by nonresponse in the various subpopulations [48].

Ethics Approval

Participation in the study was completely voluntary. Ethical approval was obtained from the Ethics Committee of the Finnish Institute for Health and Welfare (THL/637/6.02.01/2017).

Measurements

Dependent Variables

The *nonuse of My Kanta* was evaluated with the question "Have you used My Kanta in the past 12 months?" Respondents were asked to respond (1) *no* or (2) *yes*. For the analyses, the measure was binary-coded (0=user, 1=nonuser), and the users of My Kanta were set as the reference group.

The *dissatisfaction with My Kanta* was evaluated with a question concerning satisfaction: "If you have used the service, assess the quality of the service using a school grade (4-10)." In the Finnish education system, grades 8-10 represent grades from good to excellent and grades 4-7 from fail to satisfactory [49]. For the analyses, the measure was binary-coded (0=respondent was satisfied [grades 8-10] and 1=respondent was dissatisfied [grades 4-7] with the service), and the satisfied users of My Kanta were set as the reference group. Because the research interest was in respondents who were less satisfied with the service, respondents who gave an assessment of grade 7 were included in the group of dissatisfied users. This decision was also made based on substantive judgment to even the distribution [50] between satisfied and dissatisfied users, since only a small number of respondents had selected a grade from 4 to 6.

Independent Variables

Independent variables included characteristics concerning (1) sociodemographic background, (2) health and the use of health care services, (3) experiences of guidance concerning electronic services, and (4) digital skills and attitudes. All the used variables are presented in Multimedia Appendix 1.

Sociodemographic Characteristics

The respondents' sociodemographic characteristics included their age, gender, education, and degree of urbanization. Age was used as a categorical variable in the descriptive statistics and as a continuous variable in all analyses. The *degree of urbanization* was determined according to the municipal classification and divided into 3 categories according to the proportion of people living in urban settlements and the population of the largest urban settlement: urban, semiurban, and rural municipalities [51]. Because of age-related differences in education, the respondents' *educational level* was first divided into 10-year age groups by gender. Each group was divided into 3 categories based on their years of education, with approximately one-third of the respondents in each category: low, median, and high. Hence, the education-level variable had hardly any interaction with age and gender.

Health and the Use of Health Care Services

Variables concerning the respondents' health and the use of health care services included self-rated health, long-term illness, and the use of health care services. *Self-rated health* was evaluated with a widely used, single-item measure of self-perceived health status. A subjective assessment of own health has been reported as a more sensitive measure in health monitoring than external measures of health, since it includes biological, psychological, and social dimensions. [52]. A scale from good to poor was used to evaluate the present state of the respondents' health. In the analyses, the options (1) *good* and (2) *fairly good* were combined to represent good health, and the remaining options represented average or poor health. *Long-term illness* was binary-coded as (1) *yes* and (2) *no*. The *use of health care services* was binary-coded according to the number of annual outpatient appointments with a physician; 8 or more annual appointments were considered a high use of health care services, and less than 8 were counted as low or average use [53].

Experiences of Guidance Concerning Electronic Services

Variables concerning the experiences of guidance concerning electronic services included referrals to electronic services and the need for guidance on how to use online social or health care services. The *referral to electronic services* was evaluated with the question "If you have used social or health care services in the traditional way (paper, visit, or call) in the past 12 months, were you referred to electronic services (eg, My Kanta)?" For the analyses, option (1) *yes, I was referred* represented the respondents who were referred to electronic health care services. Option (2) was for those who were not referred to electronic health care services.

The *need for guidance* on using online social and health care services was evaluated with the statement "I need help with using online social and health care services." In the analyses, the options (1) *completely agree* and (2) *somewhat agree* were combined as (1) *yes* and the remaining options as (2) *no* or *no opinion*.

Variables Related to Digital Skills and Attitudes

Variables related to digital skills and attitudes included digital skills, perceived benefits of electronic social and health care services, and security concerns. *Digital skills* were evaluated with 6 validated statements [54]. Based on pilot testing, 2 (33%) of the statements were transferred into positive statements [37]. A 5-point Likert scale was used to answer the statements (1=completely agree to 5=strongly disagree). Cronbach α for the statements was .86. In the analyses, a mean variable ranging from 1 to 5 was calculated for each respondent, and the measure was binary-coded to indicate (1) *good skills* (mean≤2.5) and (2) *poor skills* (mean>2.6). The same coding has previously been used in national research [37].

The *perceived benefits* of electronic social and health care services were measured with 8 statements. A 5-point Likert scale was used to answer the statements (1=completely agree to 5=strongly disagree). Cronbach α for the statements was .91. In the analyses, missing values were coded as *neither agree nor disagree*. A mean variable from 1 to 5 was calculated for each respondent, and the measure was binary-coded as (1) *beneficial* (mean≤2.5) and (2) *unbeneficial* (mean>2.6). The same coding has previously been used in national research [37].

Security concerns were evaluated with the statement "I am concerned about information security when it comes to my personal details". In the analyses, options (1) *completely agree* and (2) *somewhat agree* were combined as (1) *yes* and the remaining options as (2) *no*.

Statistical Analysis

In all statistical analyses, the IPW method [47] was applied to correct for bias by handling both differential sampling probabilities and missing data. Due to nonresponse in some items, the number of observations varied in the analyses.

Bivariate logistic regression analyses were used to examine the adjusted associations of respondent characteristics with the nonuse of My Kanta and dissatisfaction with the service (in separate analyses). First, univariate analyses, adjusted for age, gender, and education, were conducted at a time to examine the association of the dependent variable with each independent variable. Second, a multivariable model was formed, including only those independent variables with a *P* value of <.10. This cut-off for the *P* value was used for including the variables in the multivariable model, because the purpose was to identify potential independent variables rather than to test a hypothesis [55]. In the fully adjusted multivariable model, a *P* value of <.05 was considered statistically significant. Statistical methods suitable for weighted data were used, and SPSS Statistics version 27 was applied for the analyses.

Results

Characteristics

The weighted majority (3330/3919, 85.0%) of the respondents had used My Kanta in the past 12 months. Most of the My Kanta users (2841/3330, 85.3%) were satisfied with the service. A minority of respondents (589/3919, 15%) had not used My Kanta in the past 12 months.

The IPW weighted characteristics of the respondents representative of the Finnish population are presented in Tables 1-4. Almost half of the respondents were aged between 35 and 59 years. Over half (n=3401, 56.4%) of the respondents were female, and the majority lived in urban regions. Over half of the respondents were not referred to electronic health care services, such as My Kanta, by a health care professional. About half of the respondents perceived electronic health care services to be beneficial. The respondents who had used My Kanta in the past 12 months were mostly satisfied with the service (mean 8.31, SE .03), whereas a minority of users (489/3330, 14.7%) were dissatisfied. Over one-third (1359/3919, 34.7%) of the respondents had also used an electronic service provided by their occupational health care provider. Of these respondents, a minority (130/1359, 9.6%) only used the service provided by their occupational health care provider and not the nationwide patient portal My Kanta.

Table 1. Sociodemographic characteristics of the weighted study sample[a].

Characteristics	Respondents, n (%)	Nonusers (N=589), n (%)	Dissatisfied users (N=489), n (%)
Age, years (N=3919)			
20-34	918 (23.4)	144 (24.5)	120 (24.5)
35-59	1773 (45.2)	286 (48.5)	234 (48.0)
60-74	1008 (25.7)	129 (21.9)	107 (21.8)
75-99	220 (5.6)	30 (5.1)	28 (5.7)
Gender (N=3919)			
Male	1641 (41.9)	301 (51.1)	221 (45.2)
Female	2278 (58.1)	288 (48.9)	268 (54.8)
Education (N=3873)			
Low	1499 (38.7)	232 (39.3)	196 (40.1)
Median	1254 (32.4)	187 (31.8)	126 (25.8)
High	1120 (28.9)	170 (28.9)	167 (34.1)
Degree of urbanization (N=3919)			
Urban	2918 (74.5)	427 (72.5)	373 (76.4)
Semiurban	536 (13.7)	78 (13.3)	64 (13.1)
Rural	465 (11.9)	84 (14.2)	52 (10.6)

[a]Inverse probability weighting (IPW)-corrected.

Table 2. Health and service use by the weighted study sample[a].

Characteristics	Respondents, n (%)	Nonusers (N=589), n (%)	Dissatisfied users (N=489), n (%)
Self-rated health (N=3893)			
Average or poor	1251 (32.1)	133 (22.6)	245 (50.2)
Good	2642 (67.9)	456 (77.4)	244 (49.8)
Long-term illness (N=3857)			
Yes	2165 (56.1)	191 (32.4)	319 (65.2)
No	1692 (43.9)	398 (67.6)	170 (34.8)
Use of health care services (N=3835)			
Low or average	3516 (91.7)	580 (98.5)	435 (89.0)
High	319 (8.3)	9 (1.5)	54 (11.0)

[a]Inverse probability weighting (IPW)-corrected.

Table 3. Experiences of guidance concerning electronic services for the weighted study sample[a].

Characteristics	Respondents, n (%)	Nonusers (N=589), n (%)	Dissatisfied users (N=489), n (%)
Referral to electronic services (N=3166)			
Yes	1386 (43.8)	154 (26.2)	216 (44.1)
No	1780 (56.2)	435 (73.8)	273 (55.9)
Need for guidance (N=3833)			
Yes	379 (9.9)	86 (14.6)	74 (15.1)
No	3454 (90.1)	503 (85.4)	415 (84.9)

[a]Inverse probability weighting (IPW)-corrected.

Table 4. Variables related to digital skills and attitudes of the weighted study sample[a].

Characteristics	Respondents, n (%)	Nonusers (N=589), n (%)	Dissatisfied users (N=489), n (%)
Digital skills (N=3897)			
Poor	199 (5.1)	48 (8.2)	25 (5.1)
Good	3698 (94.9)	541 (91.8)	464 (94.9)
Perceived benefits (N=3919)			
Yes	1840 (47.0)	252 (42.8)	181 (37.1)
No	2079 (53.0)	337 (57.2)	308 (62.9)
Security concerns (N=3823)			
Yes or N/A[b]	1923 (50.3)	323 (54.9)	312 (63.9)
No	1900 (49.7)	266 (45.1)	177 (36.1)

[a]Inverse probability weighting (IPW)-corrected.

[b]N/A: not applicable.

Associations With the Nonuse of My Kanta

Based on the results of age-, gender-, and education-adjusted univariate logistic regression analysis (Table 5), the following factors were included in the multivariable model: self-rated health, long-term illness, use of health care services, referral to electronic services, need for guidance, and digital skills.

The results of the fully adjusted logistic regression analysis regarding the nonuse of My Kanta are presented in Table 6. Male respondents were more likely to be nonusers of My Kanta compared to females. Respondents who used health care services to a low or average degree and who did not have a long-term illness were more likely to be nonusers of My Kanta compared to those who used health care services to a high degree and had a long-term illness. In addition, respondents who were not referred to electronic services, needed guidance, or had poor digital skills were over 2 times more likely to be nonusers of My Kanta compared to their counterparts.

Table 5. Results of univariate logistic regression analyses for the nonuse of My Kanta[a,b].

Univariate analyses characteristics[c]	OR[d] (95% CI)	*P* value[e]
Sociodemographic characteristics		
Age (years)	0.84 (0.61-1.16)	.17
Gender (male)	1.61 (1.21-2.12)	.001
Low educational level	1.03 (0.73-1.46)	.85
Median educational level	1.01 (0.71-1.42)	.99
High educational level	Reference	N/A[f]
Degree of urbanization		
Urban	Reference	N/A
Semiurban	1.03 (0.70-1.52)	.87
Rural	1.34 (0.93-1.92)	.12
Health and service use		
Self-rated health (good)	1.75 (1.25-2.45)	.001
Long-term illness (no)	3.24 (2.41-4.36)	<.001
Use of health care services (low or average)	6.75 (2.49-18.31)	<.001
Experiences of guidance concerning electronic services		
Referral to electronic services (no)	2.43 (1.70-3.49)	<.001
Need for guidance (yes)	2.09 (1.44-3.05)	<.001
Variables related to digital skills and attitudes		
Digital skills (poor)	2.37 (1.50-3.76)	<.001
Perceived benefits (no)	1.25 (0.94-1.67)	.13
Security concerns (yes)	1.25 (0.94-1.67)	.13

[a]Inverse probability weighting (IPW)-corrected.

[b]The model included the main effect of each variable adjusted for age, gender, and education.

[c]Reference categories indicated in parentheses: gender: male vs female; self-rated health: good vs average or poor; long-term illness: no vs yes; use of health care services: low or average vs high; referral to electronic services: no vs yes; need for guidance: yes vs no; digital skills: poor vs good; perceived benefits: no vs yes; security concerns: yes vs no.

[d]OR: odds ratio.

[e]Significance level of $P<.10$.

[f]N/A: not applicable.

Table 6. Results of the fully adjusted logistic regression analysis for the nonuse of My Kanta (N=2328)[a].

Multivariable model characteristics[b]	OR[c] (95% CI)	*P* value[d]
Gender		
Female	Reference	N/A[e]
Male	1.67 (1.19–2.42)	.003
Self-rated health		
Average or poor	Reference	N/A
Good	1.48 (0.95–2.31)	.08
Long-term illness		
Yes	Reference	N/A
No	2.14 (1.48–3.10)	<.001
Use of health care services		
High	Reference	N/A
Low or average	4.66 (1.29–16.84)	.02
Referral to electronic services		
Yes	Reference	N/A
No	2.51 (1.70–3.71)	<.001
Digital skills		
Good	Reference	N/A
Poor	2.53 (1.32–4.83)	.01
Need for guidance		
No	Reference	N/A
Yes	2.26 (1.41–3.65)	<.001

[a]Inverse probability weighting (IPW)-corrected.

[b]The model included all the independent variables with a *P* value of <.10 in the univariate model adjusted for age, gender, and education.

[c]OR: odds ratio.

[d]Significance level of *P*<.05.

[e]N/A: not applicable.

Associations With Dissatisfaction With the Use of My Kanta

Based on the results of the age-, gender-, and education-adjusted univariate analyses (Table 7), the following variables were included in the multivariable model: education, self-rated health, long-term illness, need for guidance, perceived benefits, and security concerns. The results of the fully adjusted logistic regression analysis are presented in Table 8.

In the fully adjusted multivariable model, respondents who were younger, were male, and had a high level of education were more likely to be dissatisfied with My Kanta compared to their counterparts. Respondents with average or poor self-rated health were over 2 times more likely to be dissatisfied with My Kanta compared to respondents with a good perception of their own health. In addition, respondents who perceived electronic services as unbeneficial, who needed guidance, and who had security concerns were more likely to be dissatisfied with My Kanta compared to their counterparts.

Table 7. Results of univariate logistic regression analyses for dissatisfaction with My Kanta[a,b].

Univariate analyses characteristics[c]	OR[d] (95% CI)	*P* value[e]
Sociodemographic characteristics		
Age (years)	0.99 (0.99–1.01)	.33
Gender (male)	1.31 (0.95–1.82)	.09
Low educational level	Reference	N/A[f]
Median educational level	0.72 (0.49–1.06)	.09
High educational level	1.11 (0.76–1.62)	.58
Degree of urbanization		
Rural	Reference	N/A
Semiurban	1.04 (0.60–1.79)	.89
Urban	1.08 (0.72–1.61)	.72
Health and service use		
Self-rated health (average or poor)	2.45 (1.79–3.36)	<.001
Long-term illness (yes)	1.34 (0.95–1.88)	.09
Use of health care services (high)	1.29 (0.76–2.21)	.35
Experiences of guidance concerning electronic services		
Referral to electronic services (no)	1.17 (0.82–1.67)	.38
Need for guidance (yes)	2.98 (1.83–4.85)	<.001
Variables related to digital skills and attitudes		
Digital skills (poor)	1.63 (0.88–3.03)	.12
Perceived benefits (no)	1.79 (1.30–2.47)	<.001
Security concerns (yes)	2.24 (1.61–3.13)	<.001

[a]Inverse probability weighting (IPW)-corrected.

[b]The model included the main effect of each variable adjusted for age, gender, and education.

[c]Reference categories indicated in the parentheses: gender: male vs female; self-rated health: average or poor vs good; long-term illness: yes vs no; use of health care services: high vs average or low; referral to electronic services: no vs yes; need for guidance: yes vs no; digital skills: poor vs good; perceived benefits: no vs yes; security concerns: yes vs no.

[d]OR: odds ratio.

[e]Significance level of $P<.10$.

[f]N/A: not applicable.

Table 8. Results of the fully adjusted logistic regression analysis for dissatisfaction with My Kanta (N=2341)[a].

Multivariable model characteristics[b]	OR[c] (95% CI)	*P* value[d]
Age	0.99 (0.98–1.00)	.002
Gender		
Female	Reference	N/A[e]
Male	1.40 (1.00–1.94)	.05
Education		
Low	Reference	N/A
High	1.48 1.02–2.16 ()	.04
Self-rated health		
Good	Reference	N/A
Average or poor	2.10 (1.51–2.93)	<.001
Long-term illness		
No	Reference	N/A
Yes	1.02 (0.71–1.45)	.92
Perceived benefits		
Yes	Reference	N/A
No	1.52 (1.09–2.11)	.01
Security concerns		
No	Reference	N/A
Yes	1.87 (1.33–2.62)	<.001
Need for guidance		
No	Reference	N/A
Yes	2.14 (1.33–3.46)	.002

[a]Inverse probability weighting (IPW)-corrected.

[b]The model included all the independent variables with a *P* value of <.10 in the model adjusted for age, gender, and education.

[c]OR: odds ratio.

[d]Significance level of *P*<.05.

[e]N/A: not applicable.

Discussion

Principal Results

Most respondents of this nationally representative survey study had used the nationwide Finnish patient portal My Kanta in the previous 12 months and were satisfied with the service. However, more than every 10th user of health care services and the internet were nonusers of the national patient portal, and approximately the same number of users were dissatisfied with the service. Males and those in a need of guidance were more likely to be nonusers of the patient portal and dissatisfied with the service compared to women and those not needing guidance. Not having a long-term illness and low or average use of health care services were associated with the increased likelihood of nonuse of the My Kanta portal. In addition, respondents who were not referred to electronic services and who had poor digital skills were more likely to be nonusers of My Kanta compared to their counterparts. A younger age, higher education, and poor self-rated health were associated with an increased likelihood of dissatisfaction with the service. In addition, respondents who did not perceive electronic health care services to be beneficial and who had security concerns were more likely to be dissatisfied with the service compared to their counterparts.

Strengths and Limitations

Finland is 1 of the forerunners of digitalization and ranked highest in information exchange and patient-centered information processing in an international comparative study [56]. By presenting the characteristics that are associated with nonuse of and dissatisfaction with the nationwide patient portal in Finland, valuable information can be offered for national initiatives for improvement and other countries aspiring to provide their residents with access to their health care documentation. However, generalizing our findings to countries with different levels of digitalization or service system should be done with caution.

A nationally presentative sample of Finnish residents was included in the analysis. The applied IPW method has previously

been reported to improve the accuracy and generalizability of results [48]. However, the findings are based on self-reported data, with the possibility for bias, as respondents might not recall the previous happenings accurately. This could also lead to problems associated with common method variance and the inflation of the strength of relationships. In addition, some of the used independent variables are hard to explicitly measure and quantify because of their subjective nature, such as perceived benefits of electronic health care services. Although multiple factors were adjusted in the analyses, the possibility of residual confounding remains. Moreover, cross-sectional survey data do not allow drawing any confirmatory causal inferences from the results.

Since only respondents who had used the internet for transactions or for searching for information were included in the study, the results are only applicable when considering the nonuse of patient portals beyond the first-level digital divide caused by the lack of necessary devices and access to the internet. Some respondents who did not use My Kanta used additional patient portals provided by private or public providers of health care services. It is also noteworthy that some respondents might not be referred to electronic health care services, because their transactions in health care do not require further action or electronic services cannot provide support in their situation. The research concerning the users and nonusers of nationwide patient portals is sparse, and comparison is difficult as the properties provided in the portals vary, in addition to the differing patient populations and adjustments in the analyses.

Comparison With Prior Work

The use of nationwide patient portals varies by country and portal [9,10]. This study showed that the majority of Finnish residents who had used health care services and the internet in the past 12 months had used the nationwide My Kanta patient portal. The use of the patient portal increased during the COVID-19 pandemic because of the availability of coronavirus test results, easing the burden on health care services [43]. The availability of the test results has likely increased the use of My Kanta among those with no long-term illness and low use of health care services. This may also suggest that a larger proportion of this group among the respondents was reached for this study than would have been reached before the pandemic. Approximately only every 10th user of the patient portal was dissatisfied, indicating a high level of satisfaction with the service. It might be presumed that the restrictions for avoiding face-to-face encounters and fear of the infection increased the overall satisfaction with electronic health care services during the COVID-19 pandemic. However, even before the pandemic, high satisfaction with My Kanta has been reported [57-59].

The results of this study suggested that younger and more educated respondents were more likely to be dissatisfied with My Kanta compared to older and less educated respondents. The findings of this study were supported by the fact that younger generations have grown up with technology and were thus more comfortable using electronic services [60], which might result in higher expectations toward the services. Previous research on a Norwegian symptom checker [61] reported that compared to younger users, older users were more satisfied with the service because they tended to navigate it in a more superficial way without gaining awareness of the existing problems. Contradictory to the findings of this study, patients with higher education have been reported to be more satisfied with telemedicine compared to patients with lower education [62]. However, only participants with moderate or high levels of digital health literacy were recruited, which might explain the contradictory results. Among the less educated respondents, digital health literacy skills might be an important factor leading to dissatisfaction with the service [63].

This study found that respondents without any long-term illness and with low or average use of health care services are more likely to be nonusers of My Kanta, which is consistent with previous research [7,19,25,27,64]. Patients without long-term illness and lower use of health care services might have fewer needs related to health care and the use of patient portals. In addition, good self-rated health was associated with patient portal nonuse in this study, similar to studies by Moll et al [27] and Zanaboni et al [12], but only before controlling for the use of health care services and long-term illness. However, even after these adjustments, a more negative perception of patients' own health was associated with dissatisfaction with the service. Previous research has reported an almost linear association between the patients' poor perception of their own health and the number of annual outpatient visits with a physician [52]. It can be anticipated that patients with poor self-rated health have more health care needs and fewer resources and, thus, presumably higher expectations of patient portals, which may lead to dissatisfaction.

Over half of the respondents were not referred to electronic health care services by their care providers, and the nonreferred respondents were less likely to use the nationwide patient portal My Kanta compared to referred respondents. Sääskilahti et al [29] and Kong et al [5] have also reported unfamiliarity with the service among nonusers of patient portals. It is important to highlight the role of professionals and health care managers in activating and engaging patients to accept and use patient portals, because promotion is heavily associated with their use [7,29,65-69]. The promotion should be integrated into routine care processes, and individual training and support should be provided on the portal use to patients with different background demographics to help prevent the digital divide from widening [25,67,69-71]. In addition to the promotion by professionals, alternative means are also needed to increase adoption [70]. Further research is necessary to identify effective ways to integrate patient portal enrollment into clinical practice. After data collection, My Kanta received national publicity because of active marketing of the EU Digital COVID Certificate, which can be downloaded from the portal. This has further increased the public's awareness of My Kanta. In the future, it might be expected that increasingly more patients have prior knowledge and experience with the service. This might also represent a significant incentive for the public's further adoption of electronic services [4].

Although patients have prior experience in the use of electronic services and information technology tools, the ability to review

and manage medical records on patient portals should not be assumed [28]. Results of this study suggest that respondents who needed guidance in the use of electronic health care services were more likely to be nonusers of the portal or dissatisfied with the service compared to respondents without a need for guidance. It is likely that perception of a lack of ability in the use of electronic health care services or poorly designed services will alter the use and the benefits from their use, leading to dissatisfaction with the portal.

The previous literature has suggested different ways in which the use of patient portals could be promoted, including ease of entry [13,30,72], easy navigation [73,74], and reducing the required cognitive demands [73]. Electronic services should meet certain accessibility requirements [75,76] in order to support the equality of the users. Many providers of electronic health care services still struggle with meeting the demands of accessibility, impairing the position of especially patients with disabilities [77]. My Kanta fulfills these requirements at a certain level by assessing and reporting the current status of the accessibility and providing an electronic channel for feedback [78].

In addition to easy accessibility of the service, assistance on the use should be available at a low threshold. Because patients have previously been reported to seldom seek help from family members, friends, service support, or health care providers [12], electronic services should be designed to include assistance to users. New electronic introductory and teaching materials have been prepared by the system administrator of My Kanta [79], and health care professionals should guide their clients to these materials. Patients with different background demographics were involved to a limited extent in the development of these materials. Efforts to stimulate participation of especially disadvantaged patients in developmental work of patient portals [80] and educational materials should be highlighted.

Digital skills are necessary for wider patient adoption and use of patient portals [31,34,37,81,82]. The findings of this study are similar to previous research [26,30] as respondents with poor digital skills were more likely to be nonusers of My Kanta portal compared to respondents with good skills. For the patients to be able to effectively navigate the portal, their digital competence needs to be promoted [71]. The Finnish national strategy for applying information technology to health care and social welfare currently states that Finnish residents should be able to use electronic services and produce self-recorded data to promote their well-being [83]. However, by making digital skills a policy priority, the equal use of electronic health care services and patient portals could be promoted and the risk for digital divide minimized [82]. Good digital skills have also been reported to be associated with the perception that electronic services are more useful [24].

Attitudes about the usefulness, appropriateness, and potential downsides of electronic services may encourage or impede the use [31]. In this study, the respondents who did not perceive overall benefits in electronic health care services were more likely dissatisfied with My Kanta compared to respondents with a more positive attitude. Negative attitudes have previously been reported to alter patient satisfaction with the patient portal [69]. To ensure the widespread and equal use of electronic health care services, all users must experience them as beneficial [24]. Patients' perceptions of the benefits can be increased by offering them demonstrations and information about the capabilities of the patient portal [69,84]. Perceived benefits were not associated with the nonuse of the patient portal according to the results of this study.

According to this study, respondents who had security concerns were more likely to be dissatisfied with My Kanta compared to respondents who felt more secure. Similar to the results of Woods et al [30], security concerns were not associated with the use of the patient portal. Privacy, security, and confidentiality concerns regarding medical information have been identified as barriers to the use of patient portals, and some patients may feel discomfort at having their personal health information on the internet [25,66,69,85]. Attention needs to be paid to information security and identity protection, since these are critical issues and central to widespread consumer acceptance and adoption of patient portals [82]. Security concerns also complicate requests for assistance and guidance from non-health-care professionals as well as the use of patient portals on public computers, since others might be able to see sensitive medical information on the screen [69,84]. Private facilities should be promoted in libraries and other places with public computers. The COVID-19 pandemic has complicated the requests for assistance and the use of computers in public facilities because of societal restrictions. More research is needed to understand how safety concerns could be alleviated. In addition, technology users of all ages should be equipped with knowledge of online privacy and security as a new set of cyber security skills are needed in the increasingly digital society [86].

Conclusion

According to the results of this population-based cross-sectional survey study in the era of COVID-19, patients without long-term illnesses, those not referred to electronic health care services, and those in need of guidance on the use of online social and health care services seem to be more likely nonusers of the Finnish nationwide patient portal My Kanta. Moreover, poor health and security concerns seem to be associated with dissatisfaction with the service. Interventions to promote referral to electronic health care services by professionals are needed. Attention must be paid to information security of the service as well as the alleviation of the patients' privacy concerns.

Acknowledgments

This study was supported by the Strategic Research Council at the Academy of Finland (projects 327145 and 327147), the Ministry of Social Affairs and Health (project 414919001), and NordForsk (project 100477). The authors wish to thank all the respondents for their contribution.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of the used variables.
[DOCX File , 16 KB - medinform_v10i4e37500_app1.docx]

References

1. Tiirinki H, Tynkkynen L, Sovala M, Atkins S, Koivusalo M, Rautiainen P, et al. COVID-19 pandemic in Finland: preliminary analysis on health system response and economic consequences. Health Policy Technol 2020 Dec;9(4):649-662 [FREE Full text] [doi: 10.1016/j.hlpt.2020.08.005] [Medline: 32874860]
2. Ting D, 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]
3. Power K, McCrea Z, White M, Breen A, Dunleavy B, O'Donoghue S, et al. The development of an epilepsy electronic patient portal: facilitating both patient empowerment and remote clinician-patient interaction in a post-COVID-19 world. Epilepsia 2020 Sep;61(9):1894-1905 [FREE Full text] [doi: 10.1111/epi.16627] [Medline: 32668026]
4. Stanimirovic D. eHealth patient portal: becoming an indispensable public health tool in the time of Covid-19. Stud Health Technol Inform 2021;281:880-884 [FREE Full text] [doi: 10.3233/shti210305]
5. Kong Q, Riedewald D, Askari M. Factors affecting portal usage among chronically ill patients during the COVID-19 pandemic in the Netherlands: cross-sectional study. JMIR Hum Factors 2021 Jul 19;8(3):e26003 [FREE Full text] [doi: 10.2196/26003] [Medline: 34003762]
6. Osborn C, Mayberry L, Wallston K, Johnson K, Elasy TA. Understanding patient portal use: implications for medication management. J Med Internet Res 2013 Jul 03;15(7):e133 [FREE Full text] [doi: 10.2196/jmir.2589] [Medline: 23823974]
7. Irizarry T, DeVito Dabbs A, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
8. Ammenwerth E, Neyer S, Hörbst A, Mueller G, Siebert U, Schnell-Inderst P. Adult patient access to electronic health records. Cochrane Database Syst Rev 2107;6:CD012707 [FREE Full text] [doi: 10.1002/14651858.CD012707]
9. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen M, Kierkegaard P, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018 Mar;7(1):44-56 [FREE Full text] [doi: 10.1016/j.hlpt.2017.11.003]
10. Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P. Strategies for building ehealth infrastructures. In: Aanestad M, Grisot M, Hanseth O, Vassilakopoulou P, editors. Information Infrastructures within European Health Care: Working with the Installed Base. Cham: Springer International; 2017:35-51.
11. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore J, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019 May 06;21(5):e13876 [FREE Full text] [doi: 10.2196/13876]
12. Zanaboni P, Kummervold P, Sørensen T, Johansen MA. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020 Feb 07;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
13. Graham T, Ali S, Avdagovska M, Ballermann M. Effects of a web-based patient portal on patient satisfaction and missed appointment rates: survey study. J Med Internet Res 2020 May 19;22(5):e17955 [FREE Full text] [doi: 10.2196/17955] [Medline: 32427109]
14. de Lusignan S, Mold F, Sheikh A, Majeed A, Wyatt J, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic interpretative review. BMJ Open 2014 Sep 08;4(9):e006021 [FREE Full text] [doi: 10.1136/bmjopen-2014-006021] [Medline: 25200561]
15. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt J, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015 Mar 02;65(632):e141-e151 [FREE Full text] [doi: 10.3399/bjgp15x683941]
16. Otte-Trojel T, de Bont A, Rundall T, van de Klundert J. How outcomes are achieved through patient portals: a realist review. J Am Med Inform Assoc 2014;21(4):751-757 [FREE Full text] [doi: 10.1136/amiajnl-2013-002501] [Medline: 24503882]
17. Tang P, Ash J, Bates D, Overhage J, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006 Mar 01;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.m2025]
18. Wiljer D, Urowitz S, Apatu E, DeLenardo C, Eysenbach G, Harth T, Canadian Committee for Patient Accessible Health Records. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res 2008 Oct 31;10(4):e34 [FREE Full text] [doi: 10.2196/jmir.1061] [Medline: 18974036]
19. Yamin C, Emani S, Williams D, Lipsitz S, Karson A, Wald J, et al. The digital divide in adoption and use of a personal health record. Arch Intern Med 2011 Mar 28;171(6):568-574 [FREE Full text] [doi: 10.1001/archinternmed.2011.34] [Medline: 21444847]

20. Delone W, McLean E. The DeLone and McLean model of information systems success: a ten-year update. J Manag Inf Syst 2014 Dec 23;19(4):9-30 [FREE Full text] [doi: 10.1080/07421222.2003.11045748]
21. Ronda M, Dijkhorst-Oei LT, Rutten GEHM. Reasons and barriers for using a patient portal: survey among patients with diabetes mellitus. J Med Internet Res 2014 Nov 25;16(11):e263 [FREE Full text] [doi: 10.2196/jmir.3457] [Medline: 25424228]
22. Valeur H, Lie A, Moen K. Patient rationales against the use of patient-accessible electronic health records: qualitative study. J Med Internet Res 2021 May 28;23(5):e24090 [FREE Full text] [doi: 10.2196/24090] [Medline: 34047711]
23. Walker D, Hefner J, Fareed N, Huerta T, McAlearney AS. Exploring the digital divide: age and race disparities in use of an inpatient portal. Telemed J E Health 2020 May;26(5):603-613 [FREE Full text] [doi: 10.1089/tmj.2019.0065] [Medline: 31313977]
24. Heponiemi T, Jormanainen V, Leemann L, Manderbacka K, Aalto AM, Hyppönen H. Digital divide in perceived benefits of online health care and social welfare services: national cross-sectional survey study. J Med Internet Res 2020 Jul 07;22(7):e17616 [FREE Full text] [doi: 10.2196/17616] [Medline: 32673218]
25. Powell KR. Patient-perceived facilitators of and barriers to electronic portal use: a systematic review. Comput Inform Nurs 2017;35:565-573 [FREE Full text] [doi: 10.1097/cin.0000000000000377]
26. van der Vaart R, Drossaert C, Taal E, Drossaers-Bakker K, Vonkeman H, van de Laar MAFJ. Impact of patient-accessible electronic medical records in rheumatology: use, satisfaction and effects on empowerment among patients. BMC Musculoskelet Disord 2014 Mar 26;15:102 [FREE Full text] [doi: 10.1186/1471-2474-15-102] [Medline: 24673997]
27. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018 Nov 01;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
28. Turvey C, Klein D, Fix G, Hogan T, Woods S, Simon S, et al. Blue Button use by patients to access and share health record information using the Department of Veterans Affairs' online patient portal. J Am Med Inform Assoc 2014;21(4):657-663 [FREE Full text] [doi: 10.1136/amiajnl-2014-002723] [Medline: 24740865]
29. Sääskilahti M, Aarnio E, Lämsä E, Ahonen R, Timonen J. Use and non-use of a nationwide patient portal: a survey among pharmacy customers. J Pharm Health Serv Res 2020;11:335-342 [FREE Full text] [doi: 10.1111/jphs.12368]
30. Woods S, Forsberg C, Schwartz E, Nazi K, Hibbard J, Houston T, et al. The association of patient factors, digital access, and online behavior on sustained patient portal use: a prospective cohort of enrolled users. J Med Internet Res 2017 Oct 17;19(10):e345 [FREE Full text] [doi: 10.2196/jmir.7895] [Medline: 29042345]
31. Helsper EJ. A corresponding fields model for the links between social and digital exclusion. Commun Theor 2012 Oct 15;22(4):403-426 [FREE Full text] [doi: 10.1111/j.1468-2885.2012.01416.x]
32. van Deursen AJ, van Dijk JA. The first-level digital divide shifts from inequalities in physical access to inequalities in material access. New Media Soc 2019 Feb;21(2):354-375 [FREE Full text] [doi: 10.1177/1461444818797082] [Medline: 30886536]
33. Graetz I, Gordon N, Fung V, Hamity C, Reed ME. The digital divide and patient portals: internet access explained differences in patient portal use for secure messaging by age, race, and income. Med Care 2016;54:772-779 [FREE Full text] [doi: 10.1097/mlr.0000000000000560]
34. Heponiemi T, Gluschkoff K, Leemann L, Manderbacka K, Aalto A, Hyppönen H. Digital inequality in Finland: access, skills and attitudes as social impact mediators. New Media Soc 2021 Jul 28:146144482110230 [FREE Full text] [doi: 10.1177/14614448211023007]
35. European Commission. DESI: Shaping Europe’s Digital Future. URL: https://digital-strategy.ec.europa.eu/en/policies/desi [accessed 2021-12-20]
36. Jormanainen V, Parhiala K, Niemi A, Erhola M, Keskimäki I, Kaila M. Half of the Finnish population accessed their own data: comprehensive access to personal health information online is a corner-stone of digital revolution in Finnish health and social care. FinJeHeW 2019 Nov 02;11(4):289-310 [FREE Full text] [doi: 10.23996/fjhw.83323]
37. Kyytsönen M, Aalto AM, Vehko T. Social and Health Care Online Service Use in 2020–2021: Experiences of the Population (English abstract; report 7/2021). Finland: Finnish Institute for Health and Welfare; 2021.
38. Jormanainen V, Reponen J. CAF and CAMM analyses on the first 10 years of national Kanta services in Finland. FinJeHeW 2020 Dec 23;12(4):302-315 [FREE Full text] [doi: 10.23996/fjhw.98548]
39. Kanta. Data Recorded in Kanta. URL: https://www.kanta.fi/en/data-in-kanta [accessed 2022-04-19]
40. Kanta. My Kanta Pages. URL: https://www.kanta.fi/en/my-kanta-pages [accessed 2022-04-19]
41. Aarnio E, Huupponen R, Martikainen J, Korhonen MJ. First insight to the Finnish nationwide electronic prescription database as a data source for pharmacoepidemiology research. Res Social Adm Pharm 2020 Apr;16(4):553-559 [FREE Full text] [doi: 10.1016/j.sapharm.2019.06.012] [Medline: 31253500]
42. Jormanainen V. Large-scale implementation and adoption of the Finnish national Kanta services in 2010–2017: a prospective, longitudinal, indicator-based study. FinJeHeW 2018 Dec 04;10(4):381-395 [FREE Full text] [doi: 10.23996/fjhw.74511]
43. Kanta. Review of Kanta Year 2020: Importance of Digital Services Highlighted. URL: https://www.kanta.fi/en/web/guest/notice/-/asset_publisher/cf6QCnduV1x6/content/katsaus-kanta-vuoteen-2020-digipalvelujen-tarkeys-korostui [accessed 2022-04-19]

44. Kanta. Statistics. URL: https://www.kanta.fi/en/statistics [accessed 2022-04-19]
45. Kanta. Acting on Behalf of Someone Else. URL: https://www.kanta.fi/en/acting-on-behalf-of-someone-else [accessed 2022-04-19]
46. Finnish Institute for Health and Welfare. National FinSote Survey. URL: https://thl.fi/en/web/thlfi-en/research-and-development/research-and-projects/national-finsote-survey [accessed 2021-11-08]
47. Seaman S, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013 Jun;22(3):278-295 [FREE Full text] [doi: 10.1177/0962280210395740] [Medline: 21220355]
48. Härkänen T, Kaikkonen R, Virtala E, Koskinen S. Inverse probability weighting and doubly robust methods in correcting the effects of non-response in the reimbursed medication and self-reported turnout estimates in the ATH survey. BMC Public Health 2014 Nov 06;14:1150 [FREE Full text] [doi: 10.1186/1471-2458-14-1150] [Medline: 25373328]
49. Scholaro. Finland Grading System. URL: https://www.scholaro.com/pro/Countries/Finland/Grading-System [accessed 2021-11-26]
50. Marsh C, Elliott J. Exploring Data: An Introduction to Data Analysis for Social Scientists, 2nd Edition. Hoboken, NJ: Wiley; 208.
51. Statistics Finland. Statistical Grouping of Municipalities. URL: https://www.stat.fi/meta/kas/til_kuntaryhmit_en.html [accessed 2021-09-30]
52. Miilunpalo S, Vuori I, Oja P, Pasanen M, Urponen H. Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. J Clin Epidemiol 1997 May;50(5):517-528 [FREE Full text] [doi: 10.1016/s0895-4356(97)00045-0]
53. Jyväsjärvi S. Frequent Attenders in Primary Health Care. A Cross-Sectional Study of Frequent Attenders' Psychosocial and Family Factors, Chronic Diseases and Reasons for Encounter in a Finnish Health Centre. University of Oulu. URL: http://jultika.oulu.fi/files/isbn9514264460.pdf [accessed 2022-04-19]
54. van Deursen AJ, Helsper E, Eynon R. Development and validation of the Internet Skills Scale (ISS). Inf Commun Soc 2015 Aug 25;19(6):804-823 [FREE Full text] [doi: 10.1080/1369118x.2015.1078834]
55. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: logistic regression. Perspect Clin Res 2017;8(3):148-151 [FREE Full text] [doi: 10.4103/picr.PICR_87_17] [Medline: 28828311]
56. Ammenwerth E, Duftschmid G, Al-Hamdan Z, Bawadi H, Cheung N, Cho K, et al. International comparison of six basic eHealth indicators across 14 countries: an eHealth benchmarking study. Methods Inf Med 2020 Dec;59(S 02):e46-e63 [FREE Full text] [doi: 10.1055/s-0040-1715796] [Medline: 33207386]
57. Lämsä E, Timonen J, Mäntyselkä P, Ahonen R. Pharmacy customers' experiences with the national online service for viewing electronic prescriptions in Finland. Int J Med Inform 2017 Jan;97:221-228 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.10.014] [Medline: 27919380]
58. Sääskilahti M, Ahonen R, Timonen J. Pharmacy customers' experiences of use, usability, and satisfaction of a nationwide patient portal: survey study. J Med Internet Res 2021 Jul 16;23(7):e25368 [FREE Full text] [doi: 10.2196/25368] [Medline: 34269687]
59. Sääskilahti M, Ojanen A, Ahonen R, Timonen J. Benefits, problems, and potential improvements in a nationwide patient portal: cross-sectional survey of pharmacy customers' experiences. J Med Internet Res 2021 Nov 03;23(11):e31483 [FREE Full text] [doi: 10.2196/31483] [Medline: 34730542]
60. Hargittai E. Digital na(t)ives? Variation in internet skills and uses among members of the "Net Generation". Sociol Inq 2010;80:92-113 [FREE Full text] [doi: 10.1111/j.1475-682x.2009.00317.x]
61. Marco-Ruiz L, Bønes E, de la Asunción E, Gabarron E, Aviles-Solis J, Lee E, et al. Combining multivariate statistics and the think-aloud protocol to assess human-computer interaction barriers in symptom checkers. J Biomed Inform 2017 Oct;74:104-122 [FREE Full text] [doi: 10.1016/j.jbi.2017.09.002] [Medline: 28893671]
62. Dopelt K, Avni N, Haimov-Sadikov Y, Golan I, Davidovitch N. Telemedicine and eHealth literacy in the era of COVID-19: a cross-sectional study in a peripheral clinic in Israel. Int J Environ Res Public Health 2021 Sep 10;18(18):9556 [FREE Full text] [doi: 10.3390/ijerph18189556] [Medline: 34574480]
63. Kontos E, Blake K, Chou WYS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
64. Hoogenbosch B, Postma J, de Man-van Ginkel JM, Tiemessen N, van Delden JJM, van Os-Medendorp H. Use and the users of a patient portal: cross-sectional study. J Med Internet Res 2018 Sep 17;20(9):e262 [FREE Full text] [doi: 10.2196/jmir.9418] [Medline: 30224334]
65. Hörhammer I, Kujala S, Hilama P, Heponiemi T. Building primary health care personnel's support for a patient portal while alleviating eHealth-related stress: survey study. J Med Internet Res 2021 Sep 22;23(9):e28976 [FREE Full text] [doi: 10.2196/28976] [Medline: 34550087]
66. Kerns J, Krist A, Longo D, Kuzel A, Woolf SH. How patients want to engage with their personal health record: a qualitative study. BMJ Open 2013 Jul 30;3(7):e002931 [FREE Full text] [doi: 10.1136/bmjopen-2013-002931] [Medline: 23901027]
67. Krist A, Woolf S, Bello G, Sabo R, Longo D, Kashiri P, et al. Engaging primary care patients to use a patient-centered personal health record. Ann Fam Med 2014;12(5):418-426 [FREE Full text] [doi: 10.1370/afm.1691] [Medline: 25354405]

68. Kujala S, Rajalahti E, Heponiemi T, Hilama P. Health professionals' expanding ehealth competences for supporting patients' self-management. Stud Health Technol Inform 2018;247:181-185.
69. Zhao J, Song B, Anand E, Schwartz D, Panesar M, Jackson G, et al. Barriers, facilitators, and solutions to optimal patient portal and personal health record use: a systematic review of the literature. AMIA Annu Symp Proc 2018;2017:1913-1922.
70. Clarke M, Lyden E, Ma J, King K, Siahpush M(, Michaud T, et al. Sociodemographic differences and factors affecting patient portal utilization. J Racial Ethn Health Disparities 2021 Aug;8(4):879-891 [FREE Full text] [doi: 10.1007/s40615-020-00846-z] [Medline: 32839896]
71. Tieu L, Schillinger D, Sarkar U, Hoskote M, Hahn K, Ratanawongsa N, et al. Online patient websites for electronic health record access among vulnerable populations: portals to nowhere? J Am Med Inform Assoc 2017 Apr 01;24(e1):e47-e54 [FREE Full text] [doi: 10.1093/jamia/ocw098] [Medline: 27402138]
72. Eriksson-Backa K, Hirvonen N, Enwald H, Huvila I. Enablers for and barriers to using My Kanta: a focus group study of older adults' perceptions of the National Electronic Health Record in Finland. Inform Health Soc Care 2021 Dec 02;46(4):399-411 [FREE Full text] [doi: 10.1080/17538157.2021.1902331] [Medline: 33787438]
73. Kruse C, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015 Feb 10;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
74. Pyper C, Amery J, Watson M, Crook C. Patients' experiences when accessing their on-line electronic patient records in primary care. Br J Gen Pract 2004;54(498):38-43.
75. Finlex. Laki digitaalisten palvelujen tarjoamisesta 306/2019. URL: https://www.finlex.fi/fi/laki/alkup/2019/20190306 [accessed 2022-03-31]
76. W3C Web Accessibility Initiative. W3C Accessibility Standards Overview. URL: https://www.w3.org/WAI/standards-guidelines/ [accessed 2022-03-31]
77. Alajarmeh N. Evaluating the accessibility of public health websites: an exploratory cross-country study. Univers Access Inf Soc 2021 Jan 27:1-19 [FREE Full text] [doi: 10.1007/s10209-020-00788-7] [Medline: 33526996]
78. Kanta. Accessibility Statement for the kanta.fi Web Service. URL: https://www.kanta.fi/en/web/guest/accessibility-statement-for-the-kanta.fi-web-service [accessed 2022-04-19]
79. Kanta. Get to Know Your My Kanta Pages. URL: https://www.kanta.fi/en/my-kanta-pages-online-school [accessed 2022-01-19]
80. Wildenbos G, Peute L, Jaspers M. Facilitators and barriers of electronic health record patient portal adoption by older adults: a literature study. Stud Health Technol Inform 2017;235:308-312.
81. Czaja S, Zarcadoolas C, Vaughon W, Lee C, Rockoff M, Levy J. The usability of electronic personal health record systems for an underserved adult population. Hum Factors 2015 May;57(3):491-506 [FREE Full text] [doi: 10.1177/0018720814549238] [Medline: 25875437]
82. Kahn JS, Aulakh V, Bosworth A. What it takes: characteristics of the ideal personal health record. Health Aff (Millwood) 2009;28(2):369-376 [FREE Full text] [doi: 10.1377/hlthaff.28.2.369] [Medline: 19275992]
83. Ministry of Social Affairs and Health. Information to Support Well-Being and Service Renewal. eHealth and eSocial Strategy 2020. URL: https://julkaisut.valtioneuvosto.fi/handle/10024/74459 [accessed 2022-01-18]
84. Luque A, van Keken A, Winters P, Keefer M, Sanders M, Fiscella K. Barriers and facilitators of online patient portals to personal health records among persons living with HIV: formative research. JMIR Res Protoc 2013 Jan 22;2(1):e8 [FREE Full text] [doi: 10.2196/resprot.2302] [Medline: 23612564]
85. Bidmead E, Marshall A. A case study of stakeholder perceptions of patient held records: the Patients Know Best (PKB) solution. Digit Health 2016;2:2055207616668431 [FREE Full text] [doi: 10.1177/2055207616668431] [Medline: 29942567]
86. Masur PK. How online privacy literacy supports self-data protection and self-determination in the age of information. Media Commun 2020 Jun 23;8(2):258-269 [FREE Full text] [doi: 10.17645/mac.v8i2.2855]

Abbreviations

IPW: inverse probability weighting
OR: odds ratio

Edited by C Lovis; submitted 24.02.22; peer-reviewed by T Risling, J Haverinen; comments to author 27.03.22; revised version received 01.04.22; accepted 11.04.22; published 22.04.22.

Please cite as:
Kainiemi E, Vehko T, Kyytsönen M, Hörhammer I, Kujala S, Jormanainen V, Heponiemi T
The Factors Associated With Nonuse of and Dissatisfaction With the National Patient Portal in Finland in the Era of COVID-19: Population-Based Cross-sectional Survey
JMIR Med Inform 2022;10(4):e37500
URL: https://medinform.jmir.org/2022/4/e37500
doi:10.2196/37500
PMID:35404831

Original Paper

Global Research Trends in Tyrosine Kinase Inhibitors: Coword and Visualization Study

Jiming Hu[1*], PhD; Kai Xing[2*], MD; Yan Zhang[3*], PhD; Miao Liu[4*], PhD; Zhiwei Wang[2*], PhD

[1]School of Information Management, Wuhan University, Wuhan, China

[2]Department of Cardiovascular Surgery, Renmin Hospital of Wuhan University, Wuhan, China

[3]Department of Clinical Laboratory, Renmin Hospital of Wuhan University, Wuhan, China

[4]Department of Pediatrics, Renmin Hospital of Wuhan University, Wuhan, China

[*]all authors contributed equally

Corresponding Author:
Miao Liu, PhD
Department of Pediatrics
Renmin Hospital of Wuhan University
No. 9 Zhang Zhidong Road
Wuchang District
Wuhan, 430060
China
Phone: 86 13349879352
Email: liumiao915@whu.edu.cn

Abstract

Background: Tyrosine kinase inhibitors (TKIs) have achieved revolutionary results in the treatment of a wide range of tumors, and many studies on this topic continue to be published every year. Some of the published reviews provide great value for us to understand TKIs. However, there is a lack of studies on the knowledge structure, bibliometric analysis, and visualization results in TKIs research.

Objective: This paper aims to investigate the knowledge structure, hotspots, and trends of evolution of the TKIs research by co-word analysis and literature visualization and help researchers in this field to gain a comprehensive understanding of the current status and trends.

Methods: We retrieved all academic papers about TKIs published between 2016 and 2020 from the Web of Science. By counting keywords from those papers, we generated the co-word networks by extracting the co-occurrence relationships between keywords, and then segmented communities to identify the subdirections of TKIs research by calculating the network metrics of the overall and local networks. We also mapped the association network topology, including the network within and between TKIs subdirections, to reveal the association and structure among varied subdirections. Furthermore, we detected keyword bursts by combining their burst weights and durations to reveal changes in the focus of TKIs research. Finally, evolution venation and strategic diagram were generated to reveal the trends of TKIs research.

Results: We obtained 6782 unique words (total frequency 26,175) from 5584 paper titles. Finally, 296 high-frequency words were selected with a threshold of 10 after discussion, the total frequency of which accounted for 65.41% (17,120/26,175). The analysis of burst disciplines revealed a variable number of burst words of TKIs research every year, especially in 2019 and 2020, such as HER2, pyrotinib, next-generation sequencing, immunotherapy, ALK-TKI, ALK rearrangement. By network calculation, the TKIs co-word network was divided into 6 communities: C1 (non-small–cell lung cancer), C2 (targeted therapy), C3 (chronic myeloid leukemia), C4 (HER2), C5 (pharmacokinetics), and C6 (ALK). The venation diagram revealed several clear and continuous evolution trends, such as non-small–cell lung cancer venation, chronic myeloid leukemia venation, renal cell carcinoma venation, chronic lymphocytic leukemia venation. In the strategic diagram, C1 (non-small–cell lung cancer) was the core direction located in the first quadrant, C2 (targeted therapy) was exactly at the junction of the first and fourth quadrants, which meant that C2 was developing; and C3 (chronic myeloid leukemia), C4 (HER2), and C5 (pharmacokinetics) were all immature and located in the third quadrant.

Conclusions: Using co-word analysis and literature visualization, we revealed the hotspots, knowledge structure, and trends of evolution of TKIs research between 2016 and 2020. TKIs research mainly focused on targeted therapies against varied tumors, particularly against non-small–cell lung cancer. The attention on chronic myeloid leukemia and pharmacokinetics was gradually

decreasing, but the focus on HER2 and ALK was rapidly increasing. TKIs research had shown a clear development path: TKIs research was disease focused and revolved around “gene targets/targeted drugs/resistance mechanisms.” Our outcomes will provide sound and effective support to researchers, funders, policymakers, and clinicians.

(JMIR Med Inform 2022;10(4):e34548) doi:10.2196/34548

KEYWORDS

TKIs; coword analysis; literature visualization; NSCLC; targeted therapy; CML; topics distribution; HER2; pharmacokinetics

Introduction

Background

Tyrosine kinases (TKs) are a collective term for dozens of kinases encoded by multiple genes, which can phosphorylate tyrosine residues in cells [1]. Based on varied cellular localizations, the TKs family is divided into receptor tyrosine kinases (RTKs) [2] and non-RTKs [3]. RTKs consist of 20 subfamilies (eg, epidermal growth factor receptor or EGFR [4], vascular endothelial growth factor receptor or VEGFR [5]), whereas non-RTKs include 10 subfamilies such as ABL, SRC, and CSK [6]. TKs have the common activity to catalyze the transfer of γ-phosphate groups on adenosine triphosphate to the tyrosine residues of a variety of target proteins [1,3-7], and this process plays a key role in signal transduction within the cell. Abnormal activities of TKs are closely associated with proliferation, invasion, metastasis, apoptosis, and tumor angiogenesis in non-small–cell lung cancer (NSCLC) [8], chronic myeloid leukemia (CML) [2,9], and many other tumors. Therefore, TKs have become excellent targets for tumor therapy.

Tyrosine kinase inhibitors (TKIs) are a class of small-molecule compounds that can specifically inhibit TKs. They can penetrate through the cell membrane and block the signaling pathway of tumor proliferation, with some TKIs also capable of inhibiting angiogenesis [1,10]. TKIs have revolutionized the treatment of a variety of tumors [10-12]; for example, imatinib has been a typical pioneer in successfully translating oncogene research into molecular targeted therapy. Now, TKIs have developed to the fourth generation, which aims to overcome drug resistance due to T790M and C797S mutations [13]. More than 30 small-molecule TKIs have been approved for marketing by the US Food and Drug Administration (FDA), and hundreds of drug candidates are in various stages of clinical trials [13-15]. Therefore, this article aims to understand the development process of TKIs research, identify the main research directions, and analyze the potential research hotspots.

Co-word analysis is a content analysis method to study the knowledge structure and evolutionary patterns of various fields. It can facilitate researchers to identify hotspots, composition, paradigms, and evolution of a field by calculating the word pairs and co-occurrence of noun phrases in the literature [16-18]. This method has been used widely in medical bibliometric analysis, such as precision medicine [16], neonatal ischemic-hypoxic encephalopathy [19], stem cell research [20], neural stem cells [17], tumor immunotherapy [18], disaster medicine [21], medical big data [22], surgical robotics [23], epilepsy genetics [24]. We propose to use the co-word analysis and literature visualization to explore the knowledge structure, evolution trends, and associations among subtopics of TKIs research, aiming to help clinicians and scholars have a comprehensive understanding of TKIs and to give suggestions for research and usage of TKIs.

Literature Review

In recent years, targeted therapies have become a hotspot in the development of antitumor drugs with their advantages of high selectivity and low side effects [25,26]. TKIs are revolutionary targeted drugs that inhibit tumor proliferation by interfering with or inhibiting specific proteins within cancer cells, thus exerting prominent antitumor effects [1-3]. Among them, imatinib was the first targeted antitumor drug [27], which was first approved in 2001 for the treatment of BCR-ABL–positive and Philadelphia chromosome–positive CML [3,6]. And then, the first-, second-, and third-generation TKIs, represented by gefitinib, dasatinib, and osimertinib, have been validated in hundreds of clinical trials and approved for marketing [10-12,28,29].

Genetic testing has been developed rapidly. Next-generation sequencing allows for sequencing genome and exome within days and makes it possible to identify patients with druggable mutations quickly and precisely [30]. Meanwhile, multidisciplinary collaboration between pharmacology and clinical science has brought a leap forward in basic research and clinical applications of TKIs. First, tumor-targeted therapies are the most established area for TKIs, especially in the treatment of lung cancer [12,15,31,32] and leukemia [10,11,28,29]. TKIs have improved the quality of life and extended survival in patients with advanced NSCLCs [33]. Imatinib and gefitinib have become first-line drugs due to their outstanding clinical efficacy in patients with BCR-ABL–positive CML [10,11,28,29]. Second, clinical trials of various drugs targeting HER2 and ALK (eg, trastuzumab [34,35], palivizumab [36], ceritinib [37]) have manifested excellent effects. Third, pharmacokinetics is another focus of TKIs research. Optimization and selectivity study is an important direction for continuing clinical trials after the launch of many TKIs. Besides, individualized blood concentration monitoring is important for patients with poor efficacy or severe side effects [38].

Previous Efforts

In recent years, TKIs have been widely used for tumor-targeted therapies. Numerous research efforts helped clinicians and scholars better understand TKIs and facilitated the clinical translation of study outcomes.

Based on recent reviews, the current status of TKIs research is summarized as follows: First, resistance to TKIs is becoming increasingly prominent, of which genetic mutations (eg, T790M [39,40], C797S [13], D761Y [41], L747S [42]) are the main

cause. It has become essential to find new molecular mechanisms underlying resistance to TKIs and to establish individualized dosing regimens. Second, the application of drugs such as erlotinib [43], osimertinib [15,33], and gefitinib [44] has gradually matured and occupied an important position in the treatment of various tumors such as NSCLCs [11,12]. Third, drugs targeting HER2 and ALK continue to emerge, which offers new hope for solving the plague of drug resistance [45,46]. To date, hundreds of new TKIs candidates are in various stages of clinical research [47].

Rationale for the Study

Research on TKIs continues to grow to benefit more patients. However, there is still a lot of uncharted territories to explore in TKIs research. How to discover new biomarkers of TKIs? How many new applications of TKIs have been discovered? How to select TKIs with better clinical effects and fewer side effects for targeted therapy? How to overcome multidrug resistance in patients with tumors? How to individualize the use of TKIs in precision medicine? All these questions need scientific bibliometric analysis based on the results of TKIs research. The purpose of our study is to address the following questions:

1. What is the overall knowledge structure of TKIs research?
2. What are the subdirections of TKIs research and how do they interact with each other?
3. What are the evolutionary status and development trends of TKIs research in the temporal dimension?

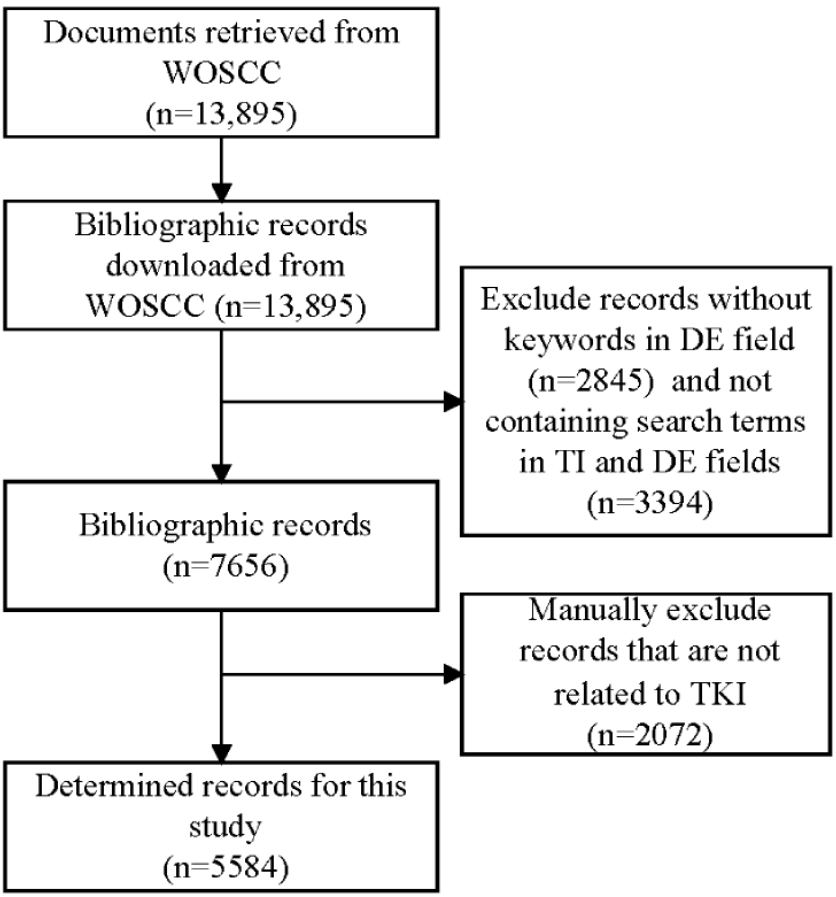
Methods

Data Collection and Processing

It is well known that Web of Science Core Collection (WOSCC) is the most extensive and comprehensive academic literature database, so we used keywords including “Tyrosine kinase inhibitor, Tyrosine kinase inhibitors, TKI, TKIs, Tyrosine kinases inhibitors, Tyrosine kinases inhibitors” in WOSCC to precisely search all studies about TKIs by limiting the period to 2016-2020 and the literature types to journal papers, reviews, and conference papers. The specific search formula was “(TS=(‘Tyrosine kinase inhibitor’ OR ‘Tyrosine kinase inhibitors’ OR ‘TKI’ OR ‘TKIs’ OR ‘Tyrosine kinases inhibitors’ OR ‘Tyrosine kinases inhibitor’)) AND LANGUAGE: (English) Refined by: DOCUMENT TYPES: (ARTICLE OR REVIEW OR PROCEEDINGS PAPER) Timespan: 2016-2020. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC.”

A total of 13,895 documents were retrieved and exported in the tab-delimited (Win) format. Next, the records containing the aforesaid search terms in the titles or keywords were retained, while those without keywords and with search terms appearing only in the abstracts were excluded [16,48]. Finally, 5584 records were obtained for the subsequent co-word network analysis (Figure 1).

Figure 1. Search procedure for documents in TKIs research. DE: descriptor; TKI: tyrosine kinase inhibitor; TI: title; WOSCC: Web of Science Core Collection.



Because there are irregularities and inconsistencies in the writing of keywords submitted by the authors themselves in WOS, it is necessary to preprocess them. First, this paper aims to depict the research status of TKIs by using other terms associated with TKIs, so TKI itself as well as the synonyms and hypernyms of TKI were removed. Keywords whose meaning is broad (eg, review, development, problem) were also removed. Second, the co-word analysis generally targets high-frequency keywords and their relationships, as keywords with very low frequencies cannot reflect the main direction of this research. Therefore, this paper (1) generated a list of keywords by frequency-descending order and (2) then defined the threshold of high-frequency keywords according to the cumulative percentage of frequencies in the list [48]. In the next step, (3) keywords with frequencies below the threshold were merged into the words with the closest meanings; besides, words with the same meaning but different forms were merged, such as "BCR-ABL TKI" to "BCR-ABL" and "epidermal growth factor receptor" to "EGFR." Finally, (4) after deduplicating the merged keywords, a new list of keyword frequencies was generated.

Network Construction and Analysis

The keywords in a paper are an accurate description of its main content, so mining keywords and their relationships can help reveal the hidden connotation of a research field [49]. If 2 words co-occur in the same connotation unit (eg, keywords in a paper), they are related or similar in connotation and have consistency in connotation expression. Their co-occurrence frequency is equal to the number of papers that contain them at the same time, and the greater the frequency, the stronger the semantic association between them [50]. By constructing co-word networks and performing structural analysis and visualization, co-word analysis can effectively reveal the underlying connotations, research structures, and even evolutionary trends of a research field [51].

In this paper, the above preprocessed data were imported into SCI2 [52] for frequency statistics and co-word network generation (.net format). Then, the .net file was imported into the network analysis tool Pajek [53] to calculate network indicators, including centralization and centrality [54], density [55], and the clustering coefficient [56], and to perform community segmentation to identify major subdirections. Centralization refers to the centripetal or consistency of the co-word network as a whole, while centrality reflects the keywords' position in the network and their ability to influence and control the network [54]. Density represents the degree of association of the network as a whole, and the stronger the association, the more mature the research field. The clustering coefficient reflects the possibility that words will cluster into classes depending on the association and its strength, and the possibility that the network will be distinctly divided into several subnetworks or subclasses. Combined with the community segmentation algorithm (Louvain) [57], the co-word network will be divided into distinctive communities, each of which represents a subdirection, with strong ties within the communities and loose ties between the communities, reflecting a greater concentration or consistency in the connotative associations of words within the community. Keywords are tightly linked within communities and loosely linked between communities, reflecting that keywords possess more focused or consistent connotations within communities.

Mapping and Visualization

To show the structure and characteristics of the TKIs research more intuitively and clearly, we visualized the topology, evolutionary venations, and development trend of the co-word network.

First, the visualization of the network topology was performed. VOSviewer [58] was used for the multilevel presentation of co-word networks and communities, including the intercommunity and intracommunity association network graphs. In the network graphs, nodes represent keywords or communities, and edges represent co-occurrence relationships between words or communities. The size of nodes and the thickness of lines are proportional to the frequency of keywords and the scale of communities, respectively, and the nodes and lines belonging to different communities are distinguished by different colors. These network diagrams visualize the importance and association relationships of keywords or communities in TKIs and help to analyze the distribution and structural characteristics of TKIs research.

Second, the visualization of evolutionary venations was performed. We divided each year's records into several communities. Then we used Cortext [59] to calculate the overlapping relationships between communities in adjacent years and connected them through "tubes." In the tube diagram, bars of different colors and sizes represent communities of different sizes, and the tubes connected by several bars represent the continuation of the research theme, which can be considered as evolutionary venations. The evolutionary trends of TKIs research over time are visualized by graphically characterizing the continuity, convergence, and divergence of communities.

Third, the visualization of the developmentary degree of the subdirections of TKIs research was performed. These research communities can be considered as subdirections of TKIs research, and each community or subdirection exhibits specific development status depending on the density and centrality. So, we drew a 2D strategic diagram based on the calculation of the density and centrality of each community. The strategic diagram took centrality, which represented the core degree of research directions in TKIs, as the horizontal axis, and density, which represented the developmental maturity of research directions, as the vertical axis, and the mean of community density and centrality as the origin. Ultimately, communities were mapped into 4 quadrants to visualize the degree of centrality and maturity of different research directions in TKIs.

Fourth, the visualization of burst words was performed. The changes in keyword frequency fluctuate significantly, with some of the words appearing in sudden bursts, reflecting the existence of distinct epochal characteristics of TKIs research. Therefore, we detected keyword bursts and combined their burst weights and durations to reveal changes in the focus of TKIs research [60].

Results

Themes Involved in TKIs Research

We extracted 10,956 unique keywords from the 5584 available paper titles, and their total frequency was 28,743 (Figure 2). After preprocessing, 6782 unique words with a total frequency of 26,175 were left. After several rounds of testing and discussion, the threshold value of high-frequency words was taken as 10 in this paper. So after merging the keywords with frequencies lower than 10 into their superordinate words, we finally obtained 296 keywords for the subsequent co-word analysis (Table 1 and Multimedia Appendix 1). These 296 keywords, whose total frequency accounted for 65.41% (17,120/26,175), can represent the mainstream of TKIs research in the past 5 years and can also reflect a strong concentration trend of TKIs research.

Burst keywords can represent important changes in TKIs research. Figure 3 shows a varying number of burst words in TKIs research each year, whose duration is expressed in terms of the length of the horizontal bar and weight in terms of the area. As can be seen from Figure 3, a variable number of emergent terms have appeared in TKIs research every year since 2016, especially in 2019 and 2020, indicating the emergence of new research themes in this field every year. The greater weight of burst words in 2020-2021 (eg, HER2, pyrotinib, next-generation sequencing, COVID-19, immunotherapy, ALK-TKI, ALK rearrangement, cell-free DNA, liquid biopsy, personalized medicine) suggests that these words in TKIs research were extensively explored by researchers in 2020.

Figure 2. Yearly number of papers and words related to tyrosine kinase inhibitor (TKI) research (2016-2020).

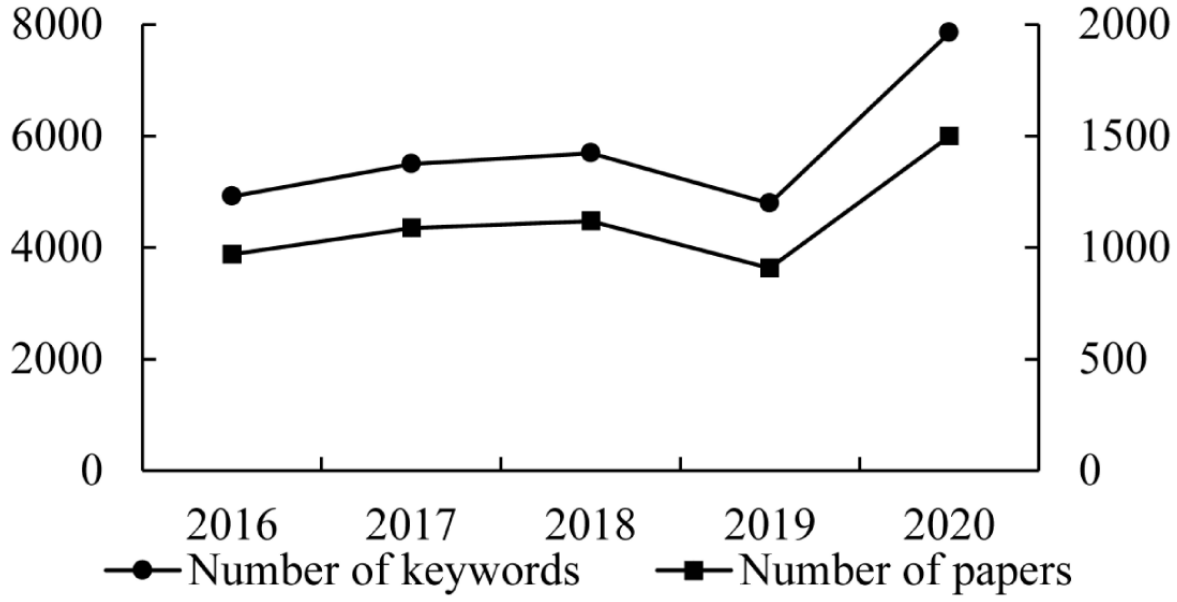
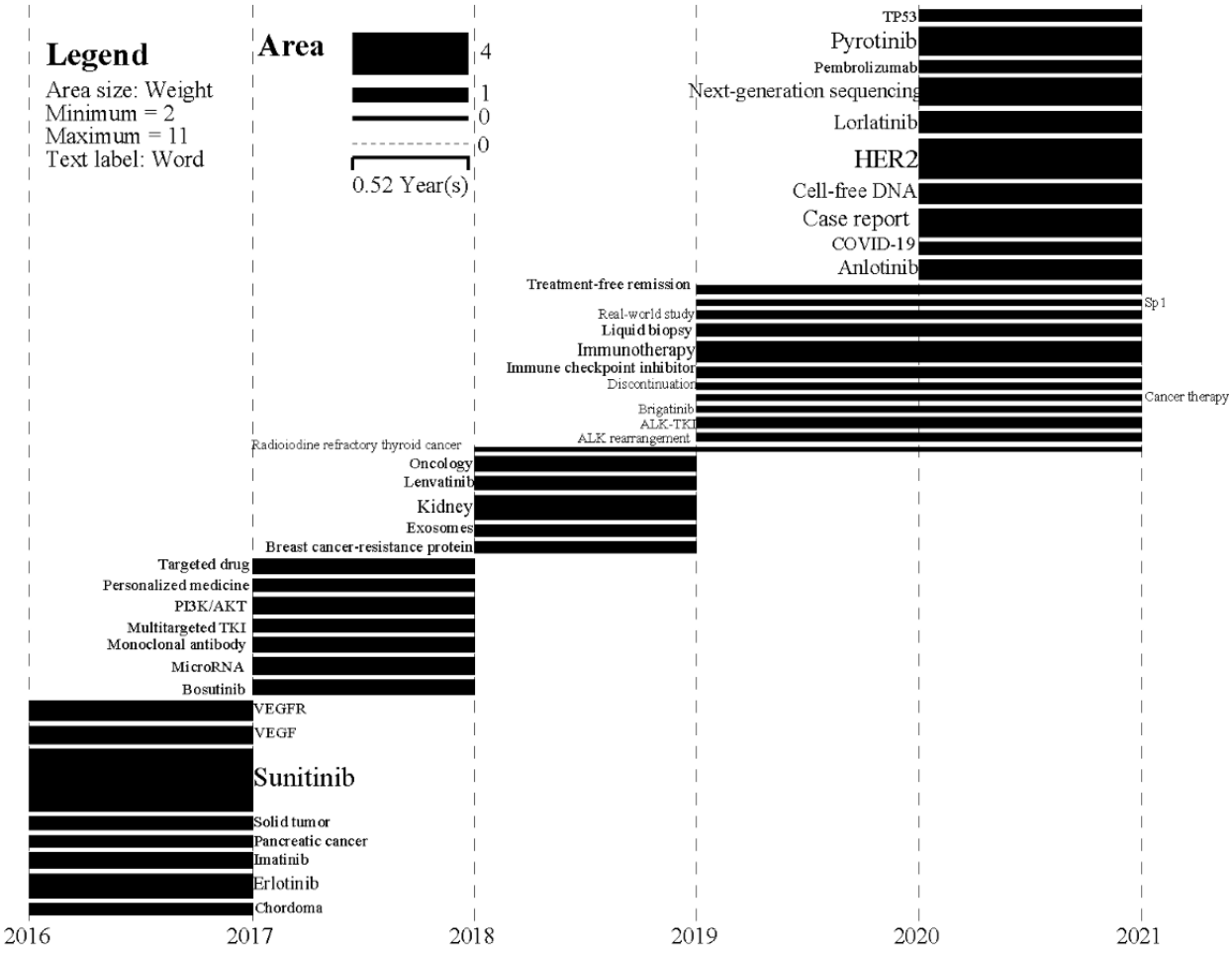


Table 1. Top 30 keywords in papers related to tyrosine kinase inhibitor (TKI) research.

Ranking	Words	Frequency
1	“Non-small cell lung cancer”	1344
2	“EGFR”	916
3	“Chronic myeloid leukemia”	586
4	“EGFR-TKI”	506
5	“EGFR mutation”	404
6	“Lung cancer”	370
7	“Erlotinib”	299
8	“Imatinib”	283
9	“Osimertinib”	261
10	“Gefitinib”	257
11	“Targeted therapy”	256
12	“Renal cell carcinoma”	227
13	“Sunitinib”	219
14	“Lung adenocarcinoma”	207
15	“Mutation”	201
16	“Resistance”	192
17	“Afatinib”	186
18	“Chemotherapy”	183
19	“Cancer”	175
20	“Dasatinib”	162
21	“Drug resistance”	159
22	“T790M”	156
23	“ALK”	156
24	“Brain metastasis”	136
25	“Tumor”	134
26	“BCR-ABL”	128
27	“Nilotinib”	121
28	“Crizotinib”	120
29	“HER2”	119
30	“Apoptosis”	113

Figure 3. Burst disciplines of TKI research from 2016 to 2021. TKI: tyrosine kinase inhibitor; VEGFR: vascular endothelial growth factor receptor.



Correlation Structure of Keywords in TKIs Research

Overview

Our analysis revealed that the co-word network consisting of the 296 keywords was exactly the maximal connected subgraph, that is, none of the high-frequency words in TKIs research is isolated, and all of them have paths associated with others, indicating that the research topics in this field form a whole that is interrelated and interact with each other.

Indicators of the Correlation Network

The overall indicators of the co-word network are shown in Table 2. The average degree of the network is 49.58, which means that a keyword in TKIs research is directly associated with 49.58 other keywords on average. These 49.58 keywords represent 16.75% of the entire network, which is a relatively small percentage, indicating that the range of intertopic associations in TKIs research is not extensive. The high degree of centralization of the network indicates a strong tendency to be centripetal or concentrated; the high closeness centralization and low betweenness centralization indicate that keywords are most directly related to each other rather than indirectly related; the high clustering coefficient indicates that keywords are likely to cluster into communities with certain words as the core. Collectively, TKIs research has clustered in certain specific subdirections in recent years, between which the distinction is obvious. However, the density of the current co-word network is not high, that is, the keywords are not closely related to each other, which indicates that TKIs research is more seriously fragmented and does not form a unified and mature research identity. We further divided the TKIs co-word network into 6 communities and calculated the module degree [61] to ensure a good division.

The network indicators of the keywords reflected their position and role in the TKIs co-word network (Table 3). Non-small cell lung cancer, EGFR, targeted therapy, lung cancer, EGFR-TKI, erlotinib, chemotherapy, cancer, sunitinib, and resistance all appear in the top 10 list of degree centralization and closeness centralization. The research topics associated with these words play an important role and have a strong influence on the whole field, while other words are likely to be clustered into a community with the above words as the core, forming a distinctive research subdirection. In addition to resistance, the above words also appear in the top 10 list of betweenness centralization, which serves as “bridges” in TKIs research, suggesting that more collaborations or synergies between TKIs research need to pass through these terms.

Table 2. The whole network indicators.

Indicators	Value
Number of nodes	296
Number of lines	7338
Average degree	49.5811
Network all degree centralization	0.5953
Network all closeness centralization	0.5265
Network betweenness centralization	0.0525
Network clustering coefficient	0.4702
Density	0.1681
Number of communities	6 (Modularity: 0.3062)

Table 3. Top 10 keywords in terms of degree, betweenness, and closeness centrality.

Ranking	Words	Degree	Words	Closeness	Words	Betweenness
1	“Non-small cell lung cancer”	224	“Non-small cell lung cancer”	0.8060	“Non-small cell lung cancer”	0.0552
2	“EGFR”	204	“EGFR”	0.7642	“EGFR”	0.0397
3	“Targeted therapy”	180	“Targeted therapy”	0.7195	“Targeted therapy”	0.0346
4	“Lung cancer”	179	“Lung cancer”	0.7178	“Chemotherapy”	0.0320
5	“EGFR-TKI”	166	“EGFR-TKI”	0.6958	“Lung cancer”	0.0291
6	“Erlotinib”	164	“Erlotinib”	0.6925	“EGFR-TKI”	0.0251
7	“Chemotherapy”	161	“Chemotherapy”	0.6876	“Erlotinib”	0.0216
8	“Cancer”	153	“Cancer”	0.6751	“Sunitinib”	0.0214
9	“Sunitinib”	151	“Sunitinib”	0.6720	“Imatinib”	0.0211
10	“Resistance”	142	“Resistance”	0.6585	“Cancer”	0.0206

Analysis of Thematic Communities

Depending on the association between keywords, TKIs are divided into 6 clusters or communities with certain important terms at their core. They are C1 (NSCLC); C2, targeted therapy; C3, CML; C4, HER2; C5, pharmacokinetics; and C6, ALK (Table 4). These 6 communities represent the subdirections of TKIs research in 2016-2020. They are each closely associated within but loosely associated with each other.

In terms of size, the subdirections of TKIs research can be divided into 3 echelons (Table 5). The first echelon includes C1 (NSCLC), which contains EGFR, EGFR-TKI, EGFR mutation, lung cancer, erlotinib, etc.; and C2 (targeted therapy), which contains renal cell carcinoma, sunitinib, chemotherapy, cancer, tumor, etc. These 2 major subdirections contain the largest number of keywords and the largest sum of frequencies, which are the main subdirections of TKIs research. The second echelon includes C3 (CML), which contains imatinib, dasatinib, BCR-ABL, nilotinib, gastrointestinal stromal tumor, etc.; C4 (HER2), which contains apoptosis, breast cancer, lapatinib, autophagy, combination therapy, etc.; and C5 (pharmacokinetics), which contains ibrutinib, metabolism, molecular docking, plasma, drug-drug interaction, etc. These 3 communities are involved in research topics that are also important for TKIs research. Keywords from these 3 communities are also important themes in TKIs research. The third echelon only contains C6 (ALK), including crizotinib, Met, RTK, ROS1, and ALK-TKI. C6 is still in its infancy, which occupies only a little weight in TKI research.

The varying centrality and density of each community further illustrate that there are sharply differentiated subdirections in TKIs research (Table 5). For example, C1 (NSCLC) and C2 (targeted therapy) have the highest centrality, which are the core subdirections in TKIs research. Besides, C1 (NSCLC) and C2 (targeted therapy) have the highest density and are the most developed subdirections in TKI research. Furthermore, the internal density of each community is higher than the density of the whole TKIs co-word network, which also indicates that each subdirection is tightly connected internally but loosely connected to each other. The centrality and density of C6 (ALK) do not have comparative value due to its small size.

Table 4. Topic communities related to tyrosine kinase inhibitor (TKI) research.

Community	Words[a]
C1-64	"Non-small cell lung cancer"; "EGFR"; "EGFR-TKI"; "EGFR mutation"; "lung cancer"; "erlotinib"; "osimertinib"; "gefitinib"; "lung adenocarcinoma"; "mutation"; "resistance"; "afatinib"; "drug resistance"; "T790M"; "brain metastasis"; "acquired resistance"; "next-generation sequencing"; "adenocarcinoma"; "T790M mutation"; "circulating tumor DNA"; "icotinib"; "liquid biopsy"; "epithelial-mesenchymal transition"; "EGFR-TKI resistance"; "sequencing"; "small cell lung cancer"; "case report"; "bevacizumab"; "gefitinib resistance"; "pemetrexed"; "squamous cell carcinoma"; "KRAS"; "epidermal growth factor"; "real-world study"; "leptomeningeal metastasis"; "advanced NSCLC"; "cost-effectiveness"; "rebiopsy"; "dacomitinib"; "IGF-1R"; "STAT3"; "droplet digital PCR"; "whole-brain radiotherapy"; "pleural effusion"; "BIM"; "uncommon mutation"; "polymorphism"; "Met amplification"; "first-line treatment"; "EGFR exon 20"; "computed tomography"; "cell-free DNA"; "TP53"; "radiosurgery"; "cisplatin; "docetaxel"; "leptomeningeal carcinomatosis"; "skin rash; "exosomes"; "cerebrospinal fluid"; "exon 19 deletion"; "exon 19"; "cetuximab"; "metformin"
C2-97	"Targeted therapy"; "renal cell carcinoma"; "sunitinib"; "chemotherapy"; "cancer"; tumor"; "metastasis"; "sorafenib"; "prognosis"; "pazopanib"; "hepatocellular carcinoma"; "lenvatinib"; "apatinib"; "immunotherapy"; "toxicity"; "metastatic renal cell carcinoma"; "radiotherapy"; "VEGF"; "progression-free survival"; "biomarker"; "angiogenesis"; "overall survival"; "adverse event"; "carcinoma"; "meta-analysis"; "sarcoma"; "VEGFR2"; "oncology"; "cabozantinib"; "renal cancer"; "VEGFR-TKI"; "PD-L1"; "VEGFR"; "clinical trial"; "axitinib"; "immune checkpoint inhibitor"; "molecular targeted therapy"; "thyroid cancer"; "FGFR"; "PDGFR"; "Phase I clinical trial"; "cardiotoxicity"; "vandetanib"; "regorafenib"; "angiogenesis inhibitor"; "PD-1"; "ovarian cancer"; "colorectal cancer"; "hypertension"; "anlotinib"; "bone metastasis"; "microRNA"; "recurrence"; "soft tissue sarcoma"; "mTOR"; "hypoxia"; "anti-angiogenesis"; "nivolumab"; "prognostic factor"; "AXL"; "RET"; "phase II clinical trial"; "everolimus"; "melanoma"; "anaplastic thyroid cancer"; "differentiated thyroid cancer"; "receptor TKI"; "clear cell renal cell carcinoma"; "medullary thyroid cancer"; "cancer therapy"; "mTOR inhibitor"; "kidney"; "adjuvant"; "tumor microenvironment"; "solid tumor"; "treatment response"; "multitargeted TKI"; "neutrophil-lymphocyte ratio"; "toceranib"; "multikinase inhibitor"; "antiangiogenic therapy"; "monoclonal antibody"; "sequential treatment"; "osteosarcoma"; "neoplasm metastasis"; "tolerability"; "esophageal cancer"; "hypothyroidism"; "circulating tumor cell"; "neoadjuvant therapy"; "Met TKI"; "PDGF"; "paclitaxel"; "neuroblastoma"; "oligoprogression"; "cervical cancer"; "pembrolizumab"
C3-46	"chronic myeloid leukemia"; "imatinib; dasatinib"; "BCR-ABL"; "nilotinib"; "gastrointestinal stromal tumor"; "acute myeloid leukemia"; "leukemia"; "molecular response"; "ponatinib"; "acute lymphoblastic leukemia"; "Philadelphia chromosome"; "TKI resistance"; "FLT3"; "kit"; "quality of life"; "adherence"; "head and neck squamous cell carcinoma"; "imatinib resistance"; "stem cell transplantation"; "leukemia stem cell"; "bosutinib"; "stem cell"; "Ph^+ALL"; "treatment-free remission"; "protein kinase inhibitor"; "pulmonary arterial hypertension"; "minimal residual disease"; "PDGFRA"; "discontinuation"; "T315I"; "adverse drug reaction"; "chronic phase"; "cytokine"; "interferon"; "c-Kit"; "midostaurin"; "single nucleotide polymorphism"; "ruxolitinib"; "BCR-ABL mutation"; "kit mutation"; "treatment discontinuation"; "rechallenge"; "Src tyrosine kinase"; "BCR-ABL TKI"; "patient-reported outcome"
C4-38	"HER2"; "apoptosis"; "breast cancer"; "lapatinib"; "autophagy"; "combination therapy"; "neratinib"; "c-Met"; "gastric cancer"; "glioblastoma"; "AKT"; proliferation"; "nanoparticles"; "ERK"; "pancreatic cancer"; "reactive oxygen species"; "cancer stem cell"; "chemoresistance"; "Src"; "hepatotoxicity"; "oxidative stress"; "cell cycle"; "pyrotinib"; "radiation"; "PI3K"; "mitochondria"; "trastuzumab"; "migration"; "NF-kappa B"; "gemcitabine"; "drug delivery"; "glioma"; "triple-negative breast cancer"; "diarrhea"; "adjuvant therapy"; "metastatic breast cancer"; "PI3K/AKT"; "invasion"
C5-37	"Pharmacokinetics"; "ibrutinib"; "metabolism"; "molecular docking"; "plasma"; "drug-drug interaction"; "P-glycoprotein"; "therapeutic drug monitoring"; "Bruton tyrosine kinase"; "BTK inhibitor"; "nintedanib"; "chronic lymphocytic leukemia"; "positron emission tomography"; "LC-MS/MS"; "personalized medicine"; "anticancer"; "lymphoma"; "breast cancer resistance protein"; "interstitial lung disease"; "spleen tyrosine kinase"; "multidrug resistance"; "molecular dynamics"; "antitumor"; "lung"; "inflammation"; "diabetes"; "anticancer drug"; "BCL-2"; "bioavailability"; "synthesis"; "human plasma"; "mantle cell lymphoma"; "idiopathic pulmonary fibrosis"; "virtual screening"; "UPLC-MS/MS"; "pulmonary fibrosis"; "blood-brain barrier"
C6-14	"ALK"; "crizotinib"; "Met"; "receptor tyrosine kinase"; "ROS1"; "ALK-TKI"; "alectinib"; "ALK rearrangement"; "immunohistochemistry"; "lorlatinib"; "BRAF"; "ceritinib"; "resistance mutation"; "brigatinib"

[a]Keywords in each community are listed in descending order of frequency.

Table 5. Indicators of 6 theme communities in tyrosine kinase inhibitor (TKI) research.

Community	Number of nodes	Number of lines	Total frequency	Average degree	Density
C1: Non-small–cell lung cancer	64	817	7106	61.8438	0.3989
C2: Targeted therapy	97	1386	4621	54.9588	0.2946
C3: Chronic myeloid leukemia	46	297	2332	36.3696	0.2807
C4: HER2	38	203	1226	44.1579	0.2812
C5: Pharmacokinetics	37	161	1230	35.8649	0.2352
C6: ALK	14	59	605	50.6429	0.6020

Visualization of the Correlation Network

The diverse correlation structure within and between each subdirection of the TKIs research is visualized in Figures 4 and 5. Figure 4 shows that there are differences in the influence of varied subdirections and that the association between these subdirections is uneven. C1 (NSCLC) and C2 (targeted therapy) have the strongest associations with the other subdirections, which reflect the strong influence of these 2 directions in TKIs research. The other four subdirections are oriented to C1 and C2, or depend on them to varying degrees. The association between C1 and C2 is significantly stronger than that between other subdirections, so C1 and C2 are the mainstream of current TKIs research. In particular, C1 is the most central and influential subdirection in the whole TKIs research, and the associations between C1 and other directions are generally strong. Isolated C6 (ALK) is not strongly associated with any other subdirection except for a closer association with C1.

Figure 5 further shows the correlation structure within each subdirection, where each term has a different location, function, and role. Each subdirection has a clear hierarchy, with the most influential terms at the core, which are important research themes, and the more distant from the core of the community, the less important the terms are. For example, EGFR, EGFR-TKI, EGFR mutation, lung cancer, and erlotinib are important research themes in C1 (NSCLC), which are closely related to the other terms or extended to other themes.

Figure 4. Correlation structure of subdirections in tyrosine kinase inhibitor (TKI) research.

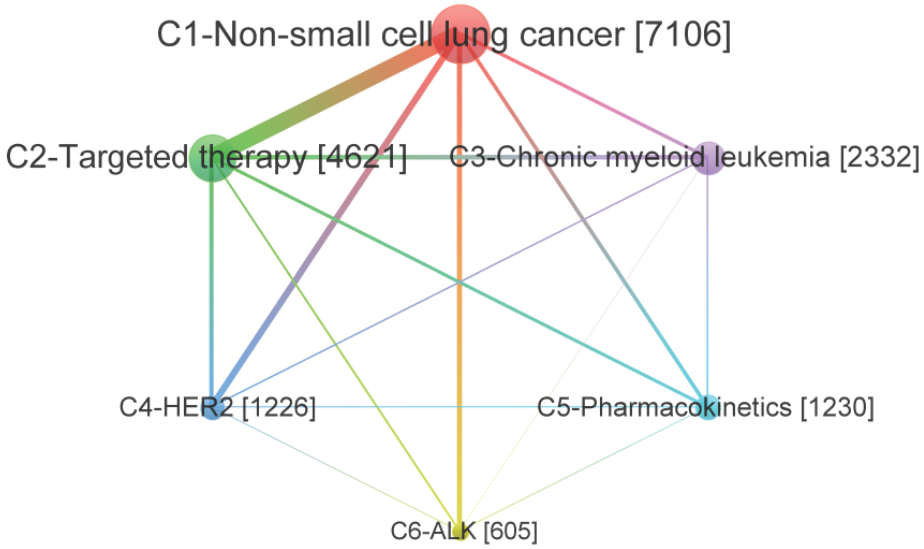
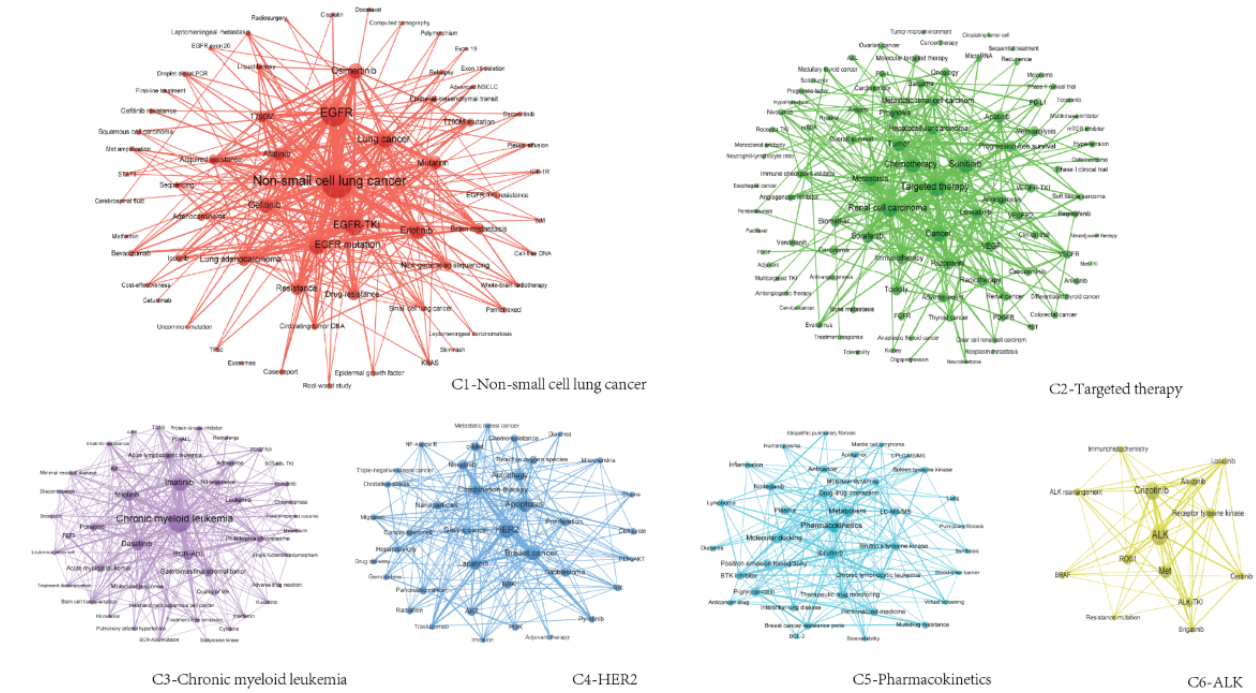


Figure 5. The internal correlation network structure of each subdirections in tyrosine kinase inhibitor (TKI) research.

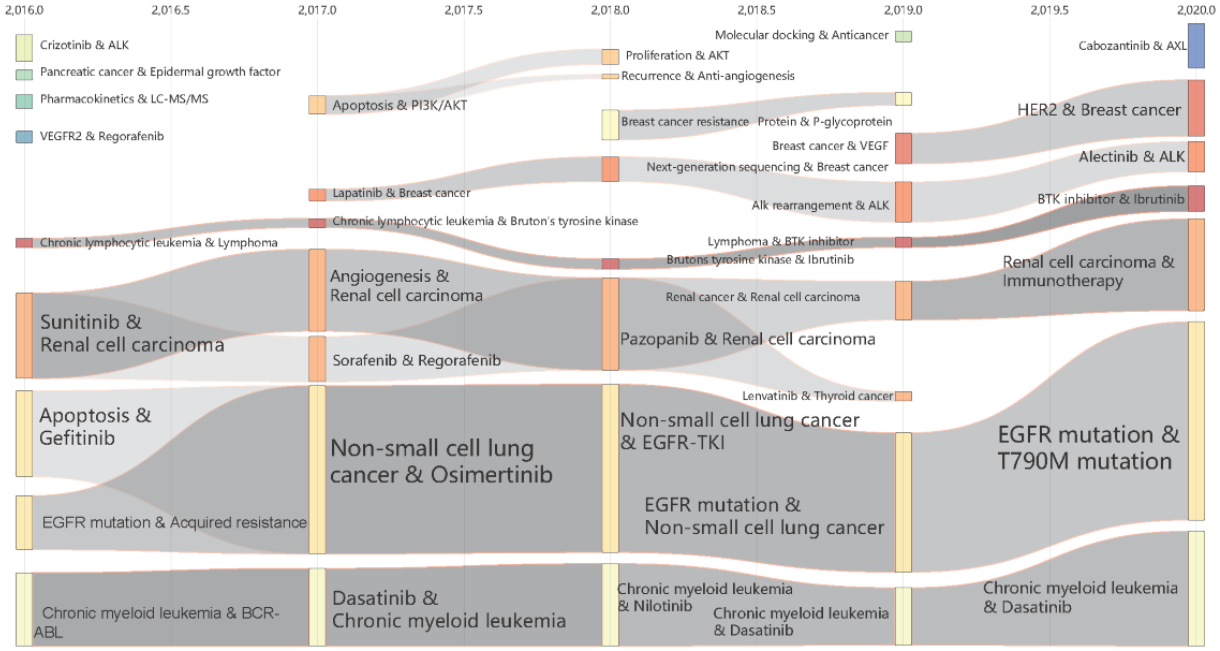


Evolution Patterns of and Trends in TKIs Research

The Evolutionary Venations of Research Themes

Figure 6 illustrates the historical evolution of TKIs research themes, which includes both the scale and the clear evolutionary venation of TKIs research themes. Overall, there are clear subdirections and good continuity of TKIs research from 2016 to 2020, but their size and distribution are uneven. In the descending order of size, several major evolutionary venations are the NSCLC venation (involving apoptosis, gefitinib, osimertinib, EGFR-TKI, T790M mutation, etc.), the CML venation (involving BCR-ABL, dasatinib, nilotinib, etc.), the renal cell carcinoma venation (involving sunitinib, angiogenesis, sorafenib, regorafenib, pazopanib, immunotherapy, etc.), the chronic lymphocytic leukemia venation (involving lymphoma, Bruton tyrosine kinase, ibrutinib, BTK inhibitor, etc.), and the lapatinib venation (involving breast cancer, next-generation sequencing, ALK rearrangement, ALK, alectinib, etc.). There are also isolated, intermittent research themes, such as crizotinib and ALK in 2016, apoptosis and PI3K/AKT in 2017, molecular docking and anticancer in 2019, and cabozantinib and AXL in 2020 for the first time as a subdirection.

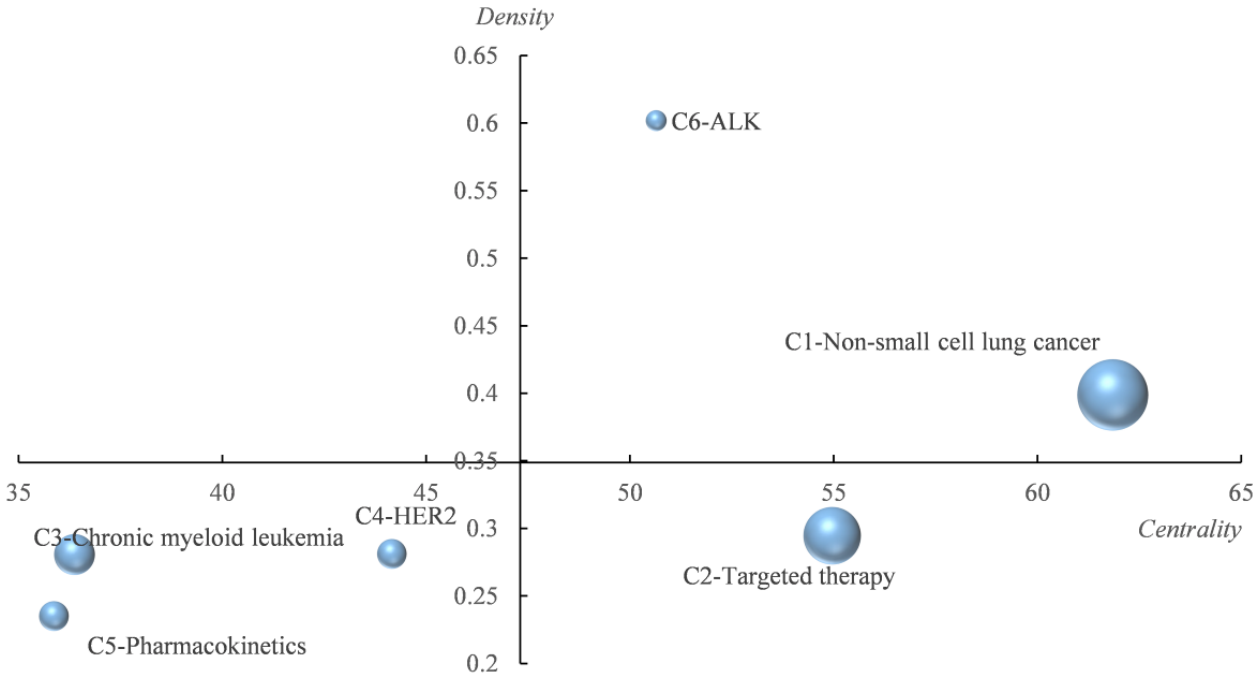
Figure 6. The evolution of themes of TKIs research over time (2016-2020). BTK: Bruton tyrosine kinase; EGFR: epidermal growth factor receptor; LC: liquid chromatography; MS/MS: tandem mass spectrometry; TKI: tyrosine kinase inhibitor; VEGFR: vascular endothelial growth factor receptor.



The Developmentary Degree of the Subdirections of TKIs Research

Based on Table 5, we drew a strategic diagram (Figure 7) to visualize the developmentary trends of the subdirections in TKIs research. The C6 community was not drawn in the strategy diagram due to its small size and noncomparable network indicators. As shown in Figure 7, we plotted the nodes of different sizes to represent the total frequencies of the varied subdirections of TKIs research and distributed them in 4 quadrants according to their density and centrality. C1 (NSCLC) is in the first quadrant due to its high density and centrality, again indicating that this community is the core direction and most developed in TKIs research. C2 (targeted therapy) is located exactly at the junction of the first and fourth quadrants, which means that C2 is also the core direction of TKIs research, but is in the process of maturing. C3 (CML), C4 (HER2), and C5 (pharmacokinetics) are all in the third quadrant, with relatively low centrality and density, which indicates that they are at the margins of TKIs research and are immature.

Figure 7. The relative development status and trends of 5 subdirections in the strategic diagram.



Discussion

Principal Findings

With the continuous progress of medical field, the targeted drugs (ie, TKIs) have made great development in basic research and have been widely used for clinical applications. Based on the results of the co-word analysis, we have a better understanding of the main research directions of TKIs and thus can accurately assess their maturity, centrality, and interactions. First, in general, TKIs research is unbalanced. The 296 words we selected from 6782 words accounted for 65.41% of the total word frequency and have a greater impact. Since the first TKIs were introduced, researchers have focused on some hot terms such as NSCLC, EGFR, CML, EGFR-TKI, EGFR mutation, erlotinib, imatinib, osimertinib, gefitinib, targeted therapy, renal cell carcinoma, resistance. These terms can be broadly classified into the following categories: clinical applications (eg, NSCLC, CML, renal cell carcinoma, lung adenocarcinoma), genetic studies (eg, EGFR, EGFR mutation, BCR-ABL), typical drugs (eg, erlotinib, imatinib, gefitinib, sunitinib, afatinib, dasatinib, osimertinib, nilotinib), and chemotherapy and drug resistance (eg, targeted therapy, resistance, chemotherapy, drug resistance). These terms not only reflect the areas of interest of researchers but also indicate the trends of TKIs research.

Based on the visualized co-word network, we found that the research topics tend to be clustered around a few keywords, eventually forming a hierarchical and relatively balanced thematic community. The thematic communities of the TKIs consist of C1 (NSCLC), C2 (targeted therapy), C3 (CML), C4 (HER2), C5 (pharmacokinetics), and C6 (ALK). There is also an imbalance between communities. First, the C1 and C2 communities are the main areas of TKIs research because of their high centrality and frequency. Among them, C1 (NSCLC), the largest thematic community, has received a lot of attention from scholars. From gene targets (eg, KRAS, EGFR, TP53, BIM) to signaling pathways (eg, IGF-1R, STAT3), from conventional chemotherapy (eg, cisplatin, docetaxel, pemetrexed) to targeted drugs (eg, erlotinib osimertinib, gefitinib), scholars have studied NSCLC in increasing depth. C2 (targeted therapy) indicates that TKIs are widely used in targeted therapies, and the application of TKIs has been extended to lung cancer [12,31,32], breast cancer [62], renal cancer [63,64], liver cancer [65], ovarian cancer [66], colorectal cancer [67], leukemia [11,28,29], thyroid cancer [68,69], cervical cancer [70], and many other tumors. Chemotherapy regimens containing TKIs were effective in reducing tumor metastasis and recurrence and improving the overall survival of patients [12,31]. The use of TKIs will be further expanded to more tumor types as more clinical trials are conducted.

Second, C3 and C5 communities have declined during the development of TKIs, both of which are in a marginal position in the strategic diagram. Among C3 (CML), imatinib is the first targeted antitumor drug that was first approved by the FDA in 2001 for patients with BCR-ABL–positive CML [10,11,28,29,71]. With the emergence of drug resistance, dasatinib [72], which targets the SRC, and nilotinib [73], which targets BCR-ABL, have been applied to patients who are resistant. Research on CML has been conducted for a long time and this field is now mature, so the application of TKIs has gradually expanded from CML to other diseases, which is leading to the gradual marginalization of the C3 community. C5 (pharmacokinetics) is an important interdisciplinary discipline related to TKIs, which is widely involved in the development process of TKIs [74,75]. However, as small-molecule drugs, the absorption, transport, distribution, and transformation of most TKIs in vivo have been clearly studied. Meanwhile, several new technologies in molecular biology (eg, molecular docking [76] and virtual high-throughput screening [77]) are used increasingly more, so C5 (pharmacokinetics) is gradually fading.

Despite the gradual decline of the C3 and C5 communities, new research areas such as C4 (HER) and C6 (ALK) have flourished in recent years. The overexpression, amplification, and mutations of HER2 have been found in a variety of tumors including breast cancer and NSCLC [45], and targeting HER2 has achieved excellent efficacy in breast cancer [62]. Although several early drugs targeting HER2 had poor efficacy in NSCLC [34,36,45], the advent of newer-generation HER2-targeting drugs such as poziotinib [78] and pyrotinib [79,80] exhibited good antitumor effects in clinical trials. Scholars are increasingly interested in targeting HER2 in NSCLC, while research on HER2 for breast cancer is relatively well established. Therefore, C4 may evolve in different directions in the future. The C6 (ALK) community is small but promising. ALK mutations, especially rearrangements, exhibit strong translational activity in NSCLC [81]. ALK-targeted agents such as alectinib [82] and brigatinib [83] have shown extraordinary efficacy in ALK-positive NSCLC and have become the first-line therapies [84]. Lorlatinib, an ALK inhibitor [85], appears as a burst word in 2020, and ALK-TKI and ALK rearrangement have a high weight from 2019 to 2020 (Figure 3), both of which indicate the rapidly rising attention on ALK. In addition, there are still several ALK-targeted drugs in development, so more literature on ALK will be published in the future and the C6 community will grow further.

We found several evolutionary venations by analyzing the evolution of themes of TKIs research over time. These highly concentrated evolutionary venations indicate scholars' continuous and steady focus on NSCLC, CML, renal cell carcinoma, chronic lymphocytic leukemia, etc. These different evolutionary lines show a clear development path: TKIs research is disease focused and revolved around "gene targets/targeted drugs/resistance mechanisms." For example, in the CML venation, investigators focused on the BRC-ABL in 2016 and on dasatinib and nilotinib in 2017-2020, which could both target BCR-ABL and overcome imatinib resistance [72,73]. In the NSCLC venation, investigators focused on EGFR genes in 2016, on EGFR-TKIs represented by osimertinib in 2017-2019, and on resistance mechanisms represented by T790M in 2020. In the renal cell carcinoma venation, investigators continued to focus on various TKIs such as sunitinib, sorafenib, regorafenib, and pazopanib, and also paid attention to gene targets such as PDGFR, FGFR, c-Kit, and VEGF, and multidrug resistance. In the chronic lymphocytic leukemia venation, investigators focused on B-cell–derived chronic lymphocytic leukemia and lymphoma in 2016, on the aberrant Bruton tyrosine kinase (BTK) from B cells in 2017, and on BTK inhibitors such as ibrutinib that can treat chronic lymphocytic leukemia and lymphoma in 2018-2020. Moreover, lapatinib, a dual EGFR/HER2 TKI [86], showed good efficacy in breast cancer, which made it an independent evolutionary venation in 2017 with continuous attention to date.

In addition to the main few evolutionary venations, we identified some isolated themes that depict the current state of TKIs research. Crizotinib gained attention as an ALK inhibitor in 2016, but its popularity declined rapidly due to its poor efficacy and the emergence of second-generation ALK inhibitors, making crizotinib and ALK an isolated topic. Circulating tumor DNA is important for efficacy assessment and prognosis analysis of tumors. The future trend in TKIs research will likely be to use next-generation sequencing or liquid biopsy technology to precisely analyze circulating tumor DNA in cell-free DNA. The PI3K/AKT pathway plays an important role in cell growth, proliferation, migration, and angiogenesis, which can be activated by RTKs. Thus, as one of the mechanisms of TKIs, the apoptosis and PI3K/AKT venation was noticed in 2017-2018. Cabozantinib is a multitarget TKI that can target 9 genes (eg, AXL, Met) [87], and cabozantinib and AXL appeared as a separate topic in 2020. Furthermore, the burst words pyrotinib and anlotinib, which accounted for a relatively large weight in 2020, are multitarget inhibitors [79,80,88]. This suggests that multitarget drugs may become an important direction for the development of TKIs and will likely receive more attention in the future. Immune checkpoint inhibitors and TKIs are important drugs for tumors. The combination of PD-L1 inhibitors (eg, pembrolizumab) and TKIs (eg, lenvatinib) in patients with malignant tumors was more effective than single drug [89], suggesting that the combination therapy was an important development direction for future tumor therapy. For well-known reasons, many patients being treated with TKIs were co-infected with SARS-CoV-2 in 2020 [90], and it was also suggested that some TKIs such as BTKs may have therapeutic effects on COVID-19 [91], which made COVID-19 a burst word in TKIs research. Because of global spread of COVID-19 in 2021, investigators' interest in TKIs for patients infected with SARS-CoV-2 would further increase.

Limitation

Our study also has some limitations: first, our search included only English literature from 2016 to 2020, while non-English literature was excluded; second, the co-word analysis did not take the quality, influence, and rigor of the literature into account, which was a common shortcoming of such papers [92-94].

Conclusions

In conclusion, we presented a visualization of TKIs research during 2016-2020 utilizing co-word analysis and the hotspots, knowledge structure, and trends of evolution revealed in our work will help researchers in the field of TKIs to gain a comprehensive understanding of the current status and trends. Based on the above results, we speculate that the general status of TKIs research is as follows: (1) NSCLC and CML are the most important clinical application areas for TKIs; (2) EGFR is the most common target gene for TKIs, and EGFR-TKIs are the most commonly used molecularly targeted TKIs, among which erlotinib, osimertinib, and gefitinib have gradually matured; (3) TKIs have become a mature field for targeted therapeutic applications, and drugs targeting HER2 and ALK have further expanded the application of TKIs; and (4) drug resistance remains a major challenge for TKIs. In a nutshell, our work remains valuable in revealing the knowledge structure and evolutionary trends of TKIs research.

Acknowledgments

This study was supported by grants from the Natural Science Foundation of Hubei Province (Grant No. 2020CFB611), the National Natural Science Foundation of China Funded Project (No. 71874125), and the Young Top-notch Talent Cultivation Program of Hubei Province.

Authors' Contributions

JH, ML, and ZW designed this research; ML and JH developed the search strategy and screened the literature; KX, ML, and JH screened and processed the keywords; JH and YZ participated in the data visualization and made the images and tables; ML, JH, and KX wrote the manuscript; KX translated the manuscript; ML and ZW reviewed and revised the manuscript; JH, KX, and YZ are common first authors. Authors ML (liumiao915@whu.edu.cn) and ZW (wangzhiwei@whu.edu.cn) are co-corresponding authors for this article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Top 100 keywords in papers related to research.
[DOCX File , 24 KB - medinform_v10i4e34548_app1.docx]

References

1. Jo DH, Kim JH, Kim JH. Targeting tyrosine kinases for treatment of ocular tumors. Arch Pharm Res 2019 Apr 23;42(4):305-318. [doi: 10.1007/s12272-018-1094-3] [Medline: 30470974]
2. Du Z, Lovly CM. Mechanisms of receptor tyrosine kinase activation in cancer. Mol Cancer 2018 Feb 19;17(1):58 [FREE Full text] [doi: 10.1186/s12943-018-0782-4] [Medline: 29455648]
3. Siveen KS, Prabhu KS, Achkar IW, Kuttikrishnan S, Shyam S, Khan AQ, et al. Role of non receptor tyrosine kinases in hematological malignances and its targeting by natural products. Mol Cancer 2018 Feb 19;17(1):31 [FREE Full text] [doi: 10.1186/s12943-018-0788-y] [Medline: 29455667]
4. An Z, Aksoy O, Zheng T, Fan Q, Weiss WA. Epidermal growth factor receptor and EGFRvIII in glioblastoma: signaling pathways and targeted therapies. Oncogene 2018 Mar 11;37(12):1561-1575 [FREE Full text] [doi: 10.1038/s41388-017-0045-7] [Medline: 29321659]
5. Ma Z, Yu Y, Badea CT, Kovacs JJ, Xiong X, Comhair S, et al. Vascular endothelial growth factor receptor 3 regulates endothelial function through β-arrestin 1. Circulation 2019 Mar 26;139(13):1629-1642 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.118.034961] [Medline: 30586762]
6. Loscocco F, Visani G, Galimberti S, Curti A, Isidori A. BCR-ABL independent mechanisms of resistance in chronic myeloid leukemia. Front Oncol 2019;9:939 [FREE Full text] [doi: 10.3389/fonc.2019.00939] [Medline: 31612105]
7. Szilveszter KP, Németh T, Mócsai A. Tyrosine kinases in autoimmune and inflammatory skin diseases. Front Immunol 2019;10:1862 [FREE Full text] [doi: 10.3389/fimmu.2019.01862] [Medline: 31447854]
8. Morris TA, Khoo C, Solomon BJ. Targeting ROS1 rearrangements in non-small cell lung cancer: crizotinib and newer generation tyrosine kinase inhibitors. Drugs 2019 Aug;79(12):1277-1286. [doi: 10.1007/s40265-019-01164-3] [Medline: 31313100]
9. Zeng P, Schmaier A. Ponatinib and other CML tyrosine kinase inhibitors in thrombosis. Int J Mol Sci 2020 Sep 08;21(18):6556 [FREE Full text] [doi: 10.3390/ijms21186556] [Medline: 32911643]
10. Radich JP, Deininger M, Abboud CN, Altman JK, Berman E, Bhatia R, et al. Chronic myeloid leukemia, version 1.2019, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2018 Sep 04;16(9):1108-1135. [doi: 10.6004/jnccn.2018.0071] [Medline: 30181422]
11. Pallera A, Altman JK, Berman E, Abboud CN, Bhatnagar B, Curtin P, et al. NCCN Guidelines insights: chronic myeloid leukemia, version 1.2017. J Natl Compr Canc Netw 2016 Dec 12;14(12):1505-1512. [doi: 10.6004/jnccn.2016.0162] [Medline: 27956535]
12. Ettinger D, Wood D, Aggarwal C, Aisner D, Akerley W, Bauman J, OCN, et al. NCCN Guidelines insights: non-small cell lung cancer, version 1.2020. J Natl Compr Canc Netw 2019 Dec;17(12):1464-1472. [doi: 10.6004/jnccn.2019.0059] [Medline: 31805526]
13. Wang S, Song Y, Liu D. EAI045: The fourth-generation EGFR inhibitor overcoming T790M and C797S resistance. Cancer Lett 2017 Jan 28;385:51-54. [doi: 10.1016/j.canlet.2016.11.008] [Medline: 27840244]
14. Tan C, Kumarakulasinghe NB, Huang Y, Ang YLE, Choo JR, Goh B, et al. Third generation EGFR TKIs: current data and future directions. Mol Cancer 2018 Feb 19;17(1):29-29 [FREE Full text] [doi: 10.1186/s12943-018-0778-0] [Medline: 29455654]

15. Remon J, Steuer C, Ramalingam S, Felip E. Osimertinib and other third-generation EGFR TKI in EGFR-mutant NSCLC patients. Ann Oncol 2018 Jan 01;29(suppl_1):i20-i27 [FREE Full text] [doi: 10.1093/annonc/mdx704] [Medline: 29462255]
16. Lyu X, Hu J, Dong W, Xu X. Intellectual structure and evolutionary trends of precision medicine research: coword analysis. JMIR Med Inform 2020 Feb 04;8(2):e11287 [FREE Full text] [doi: 10.2196/11287] [Medline: 32014844]
17. Wei W, Shi B, Guan X, Ma J, Wang Y, Liu J. Mapping theme trends and knowledge structures for human neural stem cells: a quantitative and co-word biclustering analysis for the 2013-2018 period. Neural Regen Res 2019 Oct;14(10):1823-1832 [FREE Full text] [doi: 10.4103/1673-5374.257535] [Medline: 31169201]
18. Lu K, Yu S, Yu M, Sun D, Huang Z, Xing H, et al. Bibliometric analysis of tumor immunotherapy studies. Med Sci Monit 2018 May 23;24:3405-3414 [FREE Full text] [doi: 10.12659/MSM.910724] [Medline: 29790485]
19. Huang J, Tang J, Qu Y, Zhang L, Zhou Y, Bao S, et al. Mapping the knowledge structure of neonatal hypoxic-ischemic wncephalopathy over the past decade: a co-word analysis based on keywords. J Child Neurol 2016 May 10;31(6):797-803. [doi: 10.1177/0883073815615673] [Medline: 26661482]
20. An XY, Wu QQ. Co-word analysis of the trends in stem cells field based on subject heading weighting. Scientometrics 2011 Apr 5;88(1):133-144. [doi: 10.1007/s11192-011-0374-1]
21. Wei W, Ge J, Xu S, Li M, Zhao Z, Li X, et al. Knowledge maps of disaster medicine in China based on co-word analysis. Disaster Med Public Health Prep 2019 Jun 23;13(3):405-409. [doi: 10.1017/dmp.2018.63] [Medline: 30033890]
22. Hsu W, Li J. Visualising and mapping the intellectual structure of medical big data. Journal of Information Science 2018 Jul 16;45(2):239-258. [doi: 10.1177/0165551518782824]
23. Shen L, Wang S, Dai W, Zhang Z. Detecting the interdisciplinary nature and topic hotspots of robotics in surgery: social network analysis and bibliometric study. J Med Internet Res 2019 Mar 26;21(3):e12625 [FREE Full text] [doi: 10.2196/12625] [Medline: 30912752]
24. Gan J, Cai Q, Galer P, Ma D, Chen X, Huang J, et al. Mapping the knowledge structure and trends of epilepsy genetics over the past decade: A co-word analysis based on medical subject headings terms. Medicine (Baltimore) 2019 Aug;98(32):e16782 [FREE Full text] [doi: 10.1097/MD.0000000000016782] [Medline: 31393404]
25. Lee YT, Tan YJ, Oon CE. Molecular targeted therapy: Treating cancer with specificity. Eur J Pharmacol 2018 Sep 05;834:188-196. [doi: 10.1016/j.ejphar.2018.07.034] [Medline: 30031797]
26. Krajewski KM, Braschi-Amirfarzan M, DiPiro PJ, Jagannathan JP, Shinagare AB. Molecular targeted therapy in modern oncology: imaging assessment of treatment response and toxicities. Korean J Radiol 2017;18(1):28-41 [FREE Full text] [doi: 10.3348/kjr.2017.18.1.28] [Medline: 28096716]
27. Claudiani S, Apperley JF. The argument for using imatinib in CML. Hematology Am Soc Hematol Educ Program 2018 Nov 30;2018(1):161-167 [FREE Full text] [doi: 10.1182/asheducation-2018.1.161] [Medline: 30504305]
28. O'Brien S, Radich JP, Abboud CN, Akhtari M, Altman JK, Berman E, et al. Chronic myelogenous leukemia, version 1.2015. J Natl Compr Canc Netw 2014 Nov 31;12(11):1590-1610. [doi: 10.6004/jnccn.2014.0159] [Medline: 25361806]
29. O'Brien S, Radich JP, Abboud CN, Akhtari M, Altman JK, Berman E, Ntational comprehensive cancer network. Chronic Myelogenous Leukemia, Version 1.2014. J Natl Compr Canc Netw 2013 Nov 13;11(11):1327-1340 [FREE Full text] [doi: 10.6004/jnccn.2013.0157] [Medline: 24225967]
30. Rieder D, Finotello F. Analysis of high-throughput RNA bisulfite sequencing data. Methods Mol Biol 2017;1562:143-154. [doi: 10.1007/978-1-4939-6807-7_10] [Medline: 28349459]
31. Hanna N, Johnson D, Temin S, Baker S, Brahmer J, Ellis PM, et al. Systemic therapy for stage IV non-small-cell lung cancer: American Society of Clinical Oncology clinical practice guideline update. J Clin Oncol 2017 Oct 20;35(30):3484-3515. [doi: 10.1200/JCO.2017.74.6065] [Medline: 28806116]
32. Novello S, Barlesi F, Califano R, Cufer T, Ekman S, Levra MG, ESMO Guidelines Committee. Metastatic non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2016 Sep;27(suppl 5):v1-v27. [doi: 10.1093/annonc/mdw326] [Medline: 27664245]
33. Soria J, Ohe Y, Vansteenkiste J, Reungwetwattana T, Chewaskulyong B, Lee KH, FLAURA Investigators. Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. N Engl J Med 2018 Dec 11;378(2):113-125. [doi: 10.1056/NEJMoa1713137] [Medline: 29151359]
34. Li BT, Shen R, Buonocore D, Olah ZT, Ni A, Ginsberg MS, et al. Ado-trastuzumab emtansine for patients with HER2-mutant lung cancers: results from a phase II basket trial. JCO 2018 Aug 20;36(24):2532-2537. [doi: 10.1200/jco.2018.77.9777]
35. Mazières J, Barlesi F, Filleron T, Besse B, Monnet I, Beau-Faller M, et al. Lung cancer patients with HER2 mutations treated with chemotherapy and HER2-targeted drugs: results from the European EUHER2 cohort. Ann Oncol 2016 Feb;27(2):281-286 [FREE Full text] [doi: 10.1093/annonc/mdv573] [Medline: 26598547]
36. Heymach J, Negrao M, Robichaux J, Carter B, Patel A, Altan M, et al. OA02.06 a phase II trial of poziotinib in EGFR and HER2 exon 20 mutant non-small cell lung cancer (NSCLC). Journal of Thoracic Oncology 2018 Oct;13(10):S323-S324 [FREE Full text] [doi: 10.1016/j.jtho.2018.08.243]
37. Soria J, Tan DSW, Chiari R, Wu Y, Paz-Ares L, Wolf J, et al. First-line ceritinib versus platinum-based chemotherapy in advanced ALK-rearranged non-small-cell lung cancer (ASCEND-4): a randomised, open-label, phase 3 study. Lancet 2017 Mar 04;389(10072):917-929. [doi: 10.1016/S0140-6736(17)30123-X] [Medline: 28126333]

38. Baccarani M, Druker BJ, Branford S, Kim DW, Pane F, Mongay L, et al. Long-term response to imatinib is not affected by the initial dose in patients with Philadelphia chromosome-positive chronic myeloid leukemia in chronic phase: final update from the Tyrosine Kinase Inhibitor Optimization and Selectivity (TOPS) study. Int J Hematol 2014;99(5):616-624. [doi: 10.1007/s12185-014-1566-2] [Medline: 24658916]
39. Kobayashi S, Boggon TJ, Dayaram T, Jänne PA, Kocher O, Meyerson M, et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med 2005 Feb 24;352(8):786-792. [doi: 10.1056/NEJMoa044238] [Medline: 15728811]
40. Wu S, Liu Y, Tsai M, Chang Y, Yu C, Yang P, et al. The mechanism of acquired resistance to irreversible EGFR tyrosine kinase inhibitor-afatinib in lung adenocarcinoma patients. Oncotarget 2016 Mar 15;7(11):12404-12413 [FREE Full text] [doi: 10.18632/oncotarget.7189] [Medline: 26862733]
41. Balak MN, Gong Y, Riely GJ, Somwar R, Li AR, Zakowski MF, et al. Novel D761Y and common secondary T790M mutations in epidermal growth factor receptor-mutant lung adenocarcinomas with acquired resistance to kinase inhibitors. Clin Cancer Res 2006 Nov 01;12(21):6494-6501. [doi: 10.1158/1078-0432.CCR-06-1570] [Medline: 17085664]
42. Yamaguchi F, Fukuchi K, Yamazaki Y, Takayasu H, Tazawa S, Tateno H, et al. Acquired resistance L747S mutation in an epidermal growth factor receptor-tyrosine kinase inhibitor-naïve patient: A report of three cases. Oncol Lett 2014 Feb;7(2):357-360 [FREE Full text] [doi: 10.3892/ol.2013.1705] [Medline: 24396447]
43. Steins M, Thomas M, Geißler M. Erlotinib. Recent Results Cancer Res 2018;211:1-17. [doi: 10.1007/978-3-319-91442-8_1] [Medline: 30069756]
44. Rawluk J, Waller C. Gefitinib. Recent Results Cancer Res 2018;211:235-246. [doi: 10.1007/978-3-319-91442-8_16] [Medline: 30069771]
45. Baraibar I, Mezquita L, Gil-Bazo I, Planchard D. Novel drugs targeting EGFR and HER2 exon 20 mutations in metastatic NSCLC. Crit Rev Oncol Hematol 2020 Apr;148:102906 [FREE Full text] [doi: 10.1016/j.critrevonc.2020.102906] [Medline: 32109716]
46. Du X, Shao Y, Qin H, Tai Y, Gao H. ALK-rearrangement in non-small-cell lung cancer (NSCLC). Thorac Cancer 2018 Apr 28;9(4):423-430 [FREE Full text] [doi: 10.1111/1759-7714.12613] [Medline: 29488330]
47. Dhillon S. Regorafenib: a review in metastatic colorectal cancer. Drugs 2018 Jul 26;78(11):1133-1144. [doi: 10.1007/s40265-018-0938-y] [Medline: 29943375]
48. Hu J, Zhang Y. Research patterns and trends of Recommendation System in China using co-word analysis. Information Processing & Management 2015 Jul;51(4):329-339. [doi: 10.1016/j.ipm.2015.02.002]
49. Callon M, Courtial J, Turner WA, Bauin S. From translations to problematic networks: An introduction to co-word analysis. Social Science Information 2016 Sep 03;22(2):191-235. [doi: 10.1177/053901883022002003]
50. Callon M, Courtial JP, Laville F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. Scientometrics 1991 Sep;22(1):155-205. [doi: 10.1007/BF02019280]
51. Alcaide–Muñoz L, Rodríguez–Bolívar MP, Cobo MJ, Herrera–Viedma E. Analysing the scientific evolution of e-Government using a science mapping approach. Government Information Quarterly 2017 Sep;34(3):545-555. [doi: 10.1016/j.giq.2017.05.002]
52. Börner K. Plug-and-play macroscopes. Commun. ACM 2011 Mar 01;54(3):60-69. [doi: 10.1145/1897852.1897871]
53. Doreian P, Lloyd P, Mrvar A. Partitioning large signed two-mode networks: Problems and prospects. Social Networks 2013 May;35(2):178-203. [doi: 10.1016/j.socnet.2012.01.002]
54. Freeman LC. Centrality in social networks conceptual clarification. Social Networks 1978 Jan;1(3):215-239. [doi: 10.1016/0378-8733(78)90021-7]
55. Yan E, Ding Y, Zhu Q. Mapping library and information science in China: a coauthorship network analysis. Scientometrics 2009 Jun 19;83(1):115-131. [doi: 10.1007/s11192-009-0027-9]
56. Leydesdorff L, Park HW, Wagner C. International coauthorship relations in the Social Sciences Citation Index: Is internationalization leading the Network? J Assn Inf Sci Tec 2014 Mar 10;65(10):2111-2126. [doi: 10.1002/asi.23102]
57. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/P10008]
58. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2009 Dec 31;84(2):523-538. [doi: 10.1007/s11192-009-0146-3]
59. Rosvall M, Bergstrom CT. Mapping change in large networks. PLoS One 2010 Jan 27;5(1):e8694 [FREE Full text] [doi: 10.1371/journal.pone.0008694] [Medline: 20111700]
60. Kleinberg J. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery 2003 Oct;7(4):373-397. [doi: 10.1145/775047.775061]
61. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys. Rev. E 2004 Feb 26;69(2):026113-002628. [doi: 10.1103/physreve.69.026113]
62. Wuerstlein R, Harbeck N. Neoadjuvant therapy for HER2-positive breast cancer. Rev Recent Clin Trials 2017 May 09;12(2):81-92. [doi: 10.2174/1574887112666170202165049] [Medline: 28164759]

63. Shah A, Kotecha R, Lemke E, Chandramohan A, Chaim J, Msaouel P, et al. Outcomes of patients with metastatic clear-cell renal cell carcinoma treated with second-line VEGFR-TKI after first-line immune checkpoint inhibitors. Eur J Cancer 2019 Jun;114:67-75 [FREE Full text] [doi: 10.1016/j.ejca.2019.04.003] [Medline: 31075726]
64. Tannir NM, Pal SK, Atkins MB. Second-line treatment landscape for renal cell carcinoma: a comprehensive review. Oncologist 2018 May 27;23(5):540-555 [FREE Full text] [doi: 10.1634/theoncologist.2017-0534] [Medline: 29487224]
65. Jindal A, Thadi A, Shailubhai K. Hepatocellular carcinoma: etiology and current and future drugs. J Clin Exp Hepatol 2019 Mar;9(2):221-232 [FREE Full text] [doi: 10.1016/j.jceh.2019.01.004] [Medline: 31024205]
66. Liu J, Nicum S, Reichardt P, Croitoru K, Illek B, Schmidinger M, et al. Assessment and management of diarrhea following VEGF receptor TKI treatment in patients with ovarian cancer. Gynecol Oncol 2018 Jul;150(1):173-179. [doi: 10.1016/j.ygyno.2018.03.058] [Medline: 29627080]
67. Kircher SM, Nimeiri HS, Benson AB. Targeting angiogenesis in colorectal cancer: tyrosine kinase inhibitors. Cancer J 2016;22(3):182-189. [doi: 10.1097/PPO.0000000000000192] [Medline: 27341596]
68. Ferrari SM, Centanni M, Virili C, Miccoli M, Ferrari P, Ruffilli I, et al. Sunitinib in the treatment of thyroid cancer. Curr Med Chem 2019 May 13;26(6):963-972. [doi: 10.2174/0929867324666171006165942] [Medline: 28990511]
69. Porcelli T, Sessa F, Luongo C, Salvatore D. Local ablative therapy of oligoprogressive TKI-treated thyroid cancer. J Endocrinol Invest 2019 Aug 9;42(8):871-879. [doi: 10.1007/s40618-019-1001-x] [Medline: 30628046]
70. Lv Y, Cang W, Li Q, Liao X, Zhan M, Deng H, et al. Erlotinib overcomes paclitaxel-resistant cancer stem cells by blocking the EGFR-CREB/GRβ-IL-6 axis in MUC1-positive cervical cancer. Oncogenesis 2019 Nov 26;8(12):70 [FREE Full text] [doi: 10.1038/s41389-019-0179-2] [Medline: 31772161]
71. Waller C. Imatinib mesylate. Recent Results Cancer Res 2018;212:1-27. [doi: 10.1007/978-3-319-91439-8_1] [Medline: 30069623]
72. McCafferty EH, Dhillon S, Deeks ED. Dasatinib: a review in pediatric chronic myeloid leukemia. Paediatr Drugs 2018 Dec 22;20(6):593-600. [doi: 10.1007/s40272-018-0319-8] [Medline: 30465234]
73. Sacha T, Saglio G. Nilotinib in the treatment of chronic myeloid leukemia. Future Oncol 2019 Mar;15(9):953-965. [doi: 10.2217/fon-2018-0468] [Medline: 30547682]
74. Suttorp M, Bornhäuser M, Metzler M, Millot F, Schleyer E. Pharmacology and pharmacokinetics of imatinib in pediatric patients. Expert Rev Clin Pharmacol 2018 Mar 06;11(3):219-231. [doi: 10.1080/17512433.2018.1398644] [Medline: 29076384]
75. Kucharczuk CR, Ganetsky A, Vozniak JM. Drug-drug interactions, safety, and pharmacokinetics of EGFR tyrosine kinase inhibitors for the treatment of non-small cell lung cancer. J Adv Pract Oncol 2018 Mar;9(2):189-200 [FREE Full text] [Medline: 30588353]
76. Pinzi L, Rastelli G. Molecular docking: shifting paradigms in drug discovery. Int J Mol Sci 2019 Sep 04;20(18):4331 [FREE Full text] [doi: 10.3390/ijms20184331] [Medline: 31487867]
77. Geromichalos GD, Alifieris CE, Geromichalou EG, Trafalis DT. Overview on the current status on virtual high-throughput screening and combinatorial chemistry approaches in multi-target anticancer drug discovery; Part II. J BUON 2016;21(6):1337-1358 [FREE Full text] [Medline: 28039691]
78. Robichaux J, Elamin Y, Vijayan R, Nilsson M, Hu L, He J, et al. Pan-cancer landscape and analysis of ERBB2 mutations identifies poziotinib as a clinically active inhibitor and enhancer of T-DM1 activity. Cancer Cell 2019 Oct 14;36(4):444-457.e7 [FREE Full text] [doi: 10.1016/j.ccell.2019.09.001] [Medline: 31588020]
79. Wang Y, Jiang T, Qin Z, Jiang J, Wang Q, Yang S, et al. HER2 exon 20 insertions in non-small-cell lung cancer are sensitive to the irreversible pan-HER receptor tyrosine kinase inhibitor pyrotinib. Ann Oncol 2019 Mar 01;30(3):447-455 [FREE Full text] [doi: 10.1093/annonc/mdy542] [Medline: 30596880]
80. Ma F, Ouyang Q, Li W, Jiang Z, Tong Z, Liu Y, et al. Pyrotinib or lapatinib combined with capecitabine in HER2-positive metastatic breast cancer with prior taxanes, anthracyclines, and/or trastuzumab: a randomized, phase II study. J Clin Oncol 2019 Oct 10;37(29):2610-2619. [doi: 10.1200/JCO.19.00108] [Medline: 31430226]
81. Soda M, Isobe K, Inoue A, Maemondo M, Oizumi S, Fujita Y, et al. A prospective PCR-based screening for the EML4-ALK oncogene in non-small cell lung cancer. Clin Cancer Res 2012 Aug 20;18(20):5682-5689. [doi: 10.1158/1078-0432.ccr-11-2947]
82. Herden M, Waller C. Alectinib. Recent Results Cancer Res 2018;211:247-256. [doi: 10.1007/978-3-319-91442-8_17] [Medline: 30069772]
84. Spencer SA, Riley AC, Matthew A, Di Pasqua AJ. Brigatinib: novel ALK inhibitor for non-small-cell lung cancer. Ann Pharmacother 2019 Jun 13;53(6):621-626. [doi: 10.1177/1060028018824578] [Medline: 30638036]
84. Elsayed M, Christopoulos P. Therapeutic sequencing in ALK NSCLC. Pharmaceuticals (Basel) 2021 Jan 21;14(2):80 [FREE Full text] [doi: 10.3390/ph14020080] [Medline: 33494549]
85. Shaw AT, Bauer TM, de Marinis F, Felip E, Goto Y, Liu G, CROWN Trial Investigators. First-line lorlatinib or crizotinib in advanced -positive lung cancer. N Engl J Med 2020 Nov 19;383(21):2018-2029. [doi: 10.1056/NEJMoa2027187] [Medline: 33207094]
86. Voigtlaender M, Schneider-Merck T, Trepel M. Lapatinib. Recent Results Cancer Res 2018;211:19-44. [doi: 10.1007/978-3-319-91442-8_2] [Medline: 30069757]

87. Deeks ED. Cabozantinib: a review in advanced hepatocellular carcinoma. Target Oncol 2019 Feb 14;14(1):107-113. [doi: 10.1007/s11523-019-00622-y] [Medline: 30767164]
88. Shen G, Zheng F, Ren D, Du F, Dong Q, Wang Z, et al. Anlotinib: a novel multi-targeting tyrosine kinase inhibitor in clinical development. J Hematol Oncol 2018 Sep 19;11(1):120 [FREE Full text] [doi: 10.1186/s13045-018-0664-7] [Medline: 30231931]
89. Taylor MH, Lee C, Makker V, Rasco D, Dutcus CE, Wu J, et al. Phase IB/II trial of lenvatinib plus pembrolizumab in patients with advanced renal cell carcinoma, endometrial cancer, and other selected advanced solid tumors. JCO 2020 Apr 10;38(11):1154-1163. [doi: 10.1200/jco.19.01598]
90. Başcı S, Ata N, Altuntaş F, Yiğenoğlu TN, Dal MS, Korkmaz S, Turkish Ministry of Health, Hematology Scientific Working Group. Outcome of COVID-19 in patients with chronic myeloid leukemia receiving tyrosine kinase inhibitors. J Oncol Pharm Pract 2020 Oct;26(7):1676-1682 [FREE Full text] [doi: 10.1177/1078155220953198] [Medline: 32854573]
91. McGee MC, August A, Huang W. BTK/ITK dual inhibitors: Modulating immunopathology and lymphopenia for COVID‐19 therapy. J Leukoc Biol 2020 Jul 08;109(1):49-53. [doi: 10.1002/jlb.5covr0620-306r]
92. Yao R, Ren C, Wang J, Wu G, Zhu X, Xia Z, et al. Publication trends of research on sepsis and host immune response during 1999-2019: a 20-year bibliometric analysis. Int J Biol Sci 2020;16(1):27-37 [FREE Full text] [doi: 10.7150/ijbs.37496] [Medline: 31892843]
93. Lei F, Ye J, Wang J, Xia Z. A bibliometric analysis of publications on oxycodone from 1998 to 2017. Biomed Res Int 2019 Oct 31;2019:9096201-9096209 [FREE Full text] [doi: 10.1155/2019/9096201] [Medline: 31781650]
94. Zyoud SH, Smale S, Waring WS, Sweileh WM, Al-Jabi SW. Global research trends in microbiome-gut-brain axis during 2009-2018: a bibliometric and visualized study. BMC Gastroenterol 2019 Aug 30;19(1):158 [FREE Full text] [doi: 10.1186/s12876-019-1076-z] [Medline: 31470803]

Abbreviations

ATP: adenosine triphosphate
BTK: Bruton tyrosine kinase
ctDNA: circulating tumor DNA
CML: chronic myeloid leukemia
EGFR: epidermal growth factor receptor
FDA: Food and Drug Administration
NSCLC: non-small small–cell lung cancer
RTKs: receptor tyrosine kinases
TKIs: tyrosine kinase inhibitors
TKs: tyrosine kinases
VEGFR: vascular endothelial growth factor receptor
WOSCC: Web of Science Core Collection

Edited by C Lovis; submitted 29.10.21; peer-reviewed by K Fultz Hollis, P Zarogoulidis; comments to author 02.01.22; revised version received 05.01.22; accepted 08.01.22; published 08.04.22.

Please cite as:
Hu J, Xing K, Zhang Y, Liu M, Wang Z
Global Research Trends in Tyrosine Kinase Inhibitors: Coword and Visualization Study
JMIR Med Inform 2022;10(4):e34548
URL: https://medinform.jmir.org/2022/4/e34548
doi:10.2196/34548
PMID:35072634

Original Paper

Neural Translation and Automated Recognition of ICD-10 Medical Entities From Natural Language: Model Development and Performance Assessment

Louis Falissard[1], PhD; Claire Morgand[1], MD, PhD; Walid Ghosn[1], PhD; Claire Imbaud[1], PhD; Karim Bounebache[1], PhD; Grégoire Rey[1], PhD

Centre for Epidemiology on Medical Causes of Death, Inserm, Le Kremlin Bicêtre, France

Corresponding Author:
Louis Falissard, PhD
Centre for Epidemiology on Medical Causes of Death
Inserm
80 Rue du Général Leclerc
Le Kremlin Bicêtre, 94270
France
Phone: 33 679649178
Email: louis.falissard@gmail.com

Abstract

Background: The recognition of medical entities from natural language is a ubiquitous problem in the medical field, with applications ranging from medical coding to the analysis of electronic health data for public health. It is, however, a complex task usually requiring human expert intervention, thus making it expansive and time-consuming. Recent advances in artificial intelligence, specifically the rise of deep learning methods, have enabled computers to make efficient decisions on a number of complex problems, with the notable example of neural sequence models and their powerful applications in natural language processing. However, they require a considerable amount of data to learn from, which is typically their main limiting factor. The Centre for Epidemiology on Medical Causes of Death (CépiDc) stores an exhaustive database of death certificates at the French national scale, amounting to several millions of natural language examples provided with their associated human-coded medical entities available to the machine learning practitioner.

Objective: The aim of this paper was to investigate the application of deep neural sequence models to the problem of medical entity recognition from natural language.

Methods: The investigated data set included every French death certificate from 2011 to 2016. These certificates contain information such as the subject's age, the subject's gender, and the chain of events leading to his or her death, both in French and encoded as International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) medical entities, for a total of around 3 million observations in the data set. The task of automatically recognizing ICD-10 medical entities from the French natural language–based chain of events leading to death was then formulated as a type of predictive modeling problem known as a sequence-to-sequence modeling problem. A deep neural network–based model, known as the Transformer, was then slightly adapted and fit to the data set. Its performance was then assessed on an external data set and compared to the current state-of-the-art approach. CIs for derived measurements were estimated via bootstrapping.

Results: The proposed approach resulted in an F-measure value of 0.952 (95% CI 0.946-0.957), which constitutes a significant improvement over the current state-of-the-art approach and its previously reported F-measure value of 0.825 as assessed on a comparable data set. Such an improvement makes possible a whole field of new applications, from nosologist-level automated coding to temporal harmonization of death statistics.

Conclusions: This paper shows that a deep artificial neural network can directly learn from voluminous data sets in order to identify complex relationships between natural language and medical entities, without any explicit prior knowledge. Although not entirely free from mistakes, the derived model constitutes a powerful tool for automated coding of medical entities from medical language with promising potential applications.

(JMIR Med Inform 2022;10(4):e26353) doi:10.2196/26353

KEYWORDS

machine learning; deep learning; machine translation; mortality statistics; automated medical entity recognition; ICD-10 coding

Introduction

Background

The democratization of electronic health record databases has created countless opportunities to gain precious insights in fields ranging from precision medicine to public health and epidemiology. However, these databases still present many challenges, both technical and methodological, that make their exploitation cumbersome. As an example, natural language is extensively present in some health-related databases, while being notoriously difficult to handle with traditional statistical methods and preventing most international comparisons due to language barriers. In order to counter these undesirable properties, several approaches have been devised. For instance, by encapsulating most medical entities in a standardized hierarchical tree structure, the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) [1] offers a powerful and expressive way of organizing analytics-compatible health databases. On the other hand, ICD-10 entities are significantly less intuitive for human users than natural language and require years of training and practice to handle fluently. As a consequence, the data production of classification-based medical data is usually handmade, expansive, and time-consuming. Several attempts have been made to design artificial intelligence–based systems that are able to automatically derive medical entities from natural language, some with quite promising performance [2-4]. However, all of them fall short in automating the complex production schemes inherent to medical databases, specifically in regard to their high data-quality standards.

However, recent innovations in deep artificial neural networks have achieved significant progress in natural language processing (NLP) [5,6]. In particular, their applications in the field of machine translation [7-9], fueled by increases in both data and computing power, repeatedly bring automated systems closer and closer to human-level performance. Several attempts have been made to apply these powerful techniques in an electronic health database setting, most of them with mitigated success. As an example, the current state of the art in ICD-10 entity recognition from natural language in death certificates still remains a combination of expert systems and support vector machine (SVM)–based classical machine learning [2]. Several explanations exist for this discrepancy between traditional machine translation and medical entity recognition. First, deep artificial neural network–based methods are known to require huge amounts of data for optimal performance. However, most experiments were either performed with slightly out-of-date neural architectures or with data set sizes at least an order of magnitude below what would be typically required [10]. On the other hand, the Centre for Epidemiology on Medical Causes of Death (CépiDc) has been storing French death certificates at the national scale since 2011 in both natural language and ICD-10–converted formats. The entire database amounts to just under 3 million death certificates, thus providing considerably better settings in which to investigate the potential applications of deep neural networks in medical entity recognition.

This paper formulates the process of ICD-10 entity recognition from natural language as a sequence-to-sequence (Seq2Seq) statistical modeling problem and proposes to solve it with a variation one of the state-of-the-art machine translation neural architectures, the Transformer. The Methods section focuses on describing the aforementioned statistical modeling problem and overall methodology. The Results section reports the results of the experiments that were performed on the French CépiDc data set as well as a comparison with the current state of the art. The Discussion section presents a discussion on the model's potential limitations through an error analysis and describes potential elements for improvement.

Related Work

The task of identifying ICD-10 medical entities from natural language, whether in French or in any other language, is a well-investigated problem, where several promising approaches have already been proposed. Most of these solutions were published at the Conference and Labs of the Evaluation Forum (CLEF) eHealth challenge [2,3,10], a competition held annually where teams compete to solve NLP tasks on medical textual data. For instance, the task of recognizing ICD-10 entities from death certificates, in several languages including French, have been addressed several times over the years in this competition. So far, when it comes to the task of extracting ICD-10 entities from French death certificates, the state of the art is held by the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI); they used a hybrid approach that combined data-based dictionaries for feature engineering and linear SVMs. However, nowadays, most NLP tasks are typically better handled by neural network–based architectures. These deep learning–based approaches have been applied to the problem at hand in this paper, mainly through a range of Seq2Seq architectures, as follows:

- Recurrent neural network–based encoder-decoder architectures, either with or without attention [11]
- Convolutional neural network–based encoder-decoder architectures [12,13]
- Fully attentional, although pretrained, architectures using a Bidirectional Encoder Representations from Transformers (BERT) model and transfer learning [14,15].

However, all those techniques, at least when applied to French data, failed to outperform the LIMSI's feature engineering–based approach. A possible explanation for this observation might lie in the data set that the teams were given. Indeed, their sample sizes were generally less than 200,000 observations [2]; this is usually far from enough for proper training of advanced deep learning models, as modern neural architectures in the neural translation academic literature usually train on data sets with up to tens of millions of observations [9]. This might also explain why teams using fully attentional models, which are the current state-of-the-art models in neural translation, used pretrained architectures and transfer learning

with BERT instead of training a full neural architecture end to end in a purely supervised fashion. The latter is exactly what this paper sets out to investigate and constitutes, at least to the authors' knowledge, the first attempt at training a modern, fully attentional, end-to-end trained model on a data set with a sample size compliant with the requirements of modern deep learning methods.

Methods

Ethical Considerations

The use of the mortality data investigated in this paper aligns with the mission of Inserm to produce national statistics on the medical causes of death, as listed in Article L2223-42 of the general code of local authorities (Code général des collectivités territoriales), after consulting the French National Commission for Data Protection and Liberties (Commission Nationale de l'Informatique et des Libertés).

Materials

Overview

The data set used for this study consists of every available death certificate found in the CépiDc database for the years 2011 to 2016, representing just under 3 million training examples. These documents record various types of information about their subjects, including the chain of events leading to the subject's death, written by a medical practitioner.

Causal Chain of Death

The causal chain of death constitutes the main source of information available on a death certificate in order to devise mortality statistics. It typically sums up the sequence of events that led to the subject's death, starting from immediate causes, such as cardiac arrest, and progressively expanding into the individual's past and to the underlying causes of death. The World Health Organization (WHO) provides countries with a standardized causal chain of events format, which France follows, alongside most developed countries. This WHO standard asks the medical practitioner in charge of reporting the events leading to the subject's passing to fill out a two-part form in natural language. The first part is comprised of four lines, in which the practitioner is asked to report the chain of events in inverse causal order (ie, immediate causes are reported on the first lines, and underlying causes are reported on the last lines). Although four lines are available for reporting, they do not all need to be filled. In fact, the last available lines are rarely used by the practitioner. The second part is comprised of two lines in which the practitioner is asked to report "any other significant conditions contributing to death but not related to the disease or condition causing it" [16] that the subject may have been suffering from.

In order to counter the language-dependent variability of death certificates across countries, a preprocessing step is typically applied to the causal chain of events leading to the individual's death, where each natural language–based line on the certificate is converted into a sequence of codes defined by the ICD-10 [1]. The ICD-10 is a medical classification created by the WHO that defines 14,199 medical entities (eg, diseases, signs, and symptoms) distributed over 22 chapters; entities are encoded with three or four alphanumeric decimal symbols (ie, one letter and two or three digits), 5615 of which are present in the investigated data set. Table 1 shows an example of a causal chain of events, taken from an American death certificate, in both natural language and ICD-10 formats.

Table 1. Example of a causal chain of events leading to death as written in natural language and as ICD-10 codes.

Part of form	Natural language	ICD-10[a,b] encoding
Part 1		
Line 1	Stroke in September left hemiparesis	I64 G819
Line 2	Fall scalp laceration fracture humerus	S010 W19 S423
Line 3	Coronary artery disease	I251
Line 4	Acute intracranial hemorrhage	I629
Part 2	Dementia depression hypertension	F03 F329 I10

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

[b]Some natural language lines correspond to several ICD-10 codes, whose orders matter in the overall coding process.

As previously mentioned, the process of converting the natural language–based causal chain of events leading to death into an ICD-10 format is the main focus of this paper. Consequently, the latter will be selected as the target variable and the former as the main explanatory variable for the neural network–based predictive model that will be further defined.

For reasons related to the underlying cause of death production process, the natural language–based chain of events and its ICD-10–encoded counterpart suffer from alignment errors at the line level, as shown in Table 2. Although qualitatively deemed quite rare, this misalignment phenomenon brings sufficient noise into the data set to prevent model convergence while fitting models with line-level sentence pairs.

Table 2. Death certificate from showcasing the misalignment phenomenon.

Part of form	Natural language	ICD-10[a] encoding
Part 1		
Line 1	Stroke in September left hemiparesis	I64 G819
Line 2	Fall scalp laceration fracture humerus	S010 W19 S423
Line 3	Coronary artery disease	I629[b] I251
Line 4	Acute intracranial hemorrhage[b]	N/A[c]
Part 2	Dementia depression hypertension	F03 F329 I10

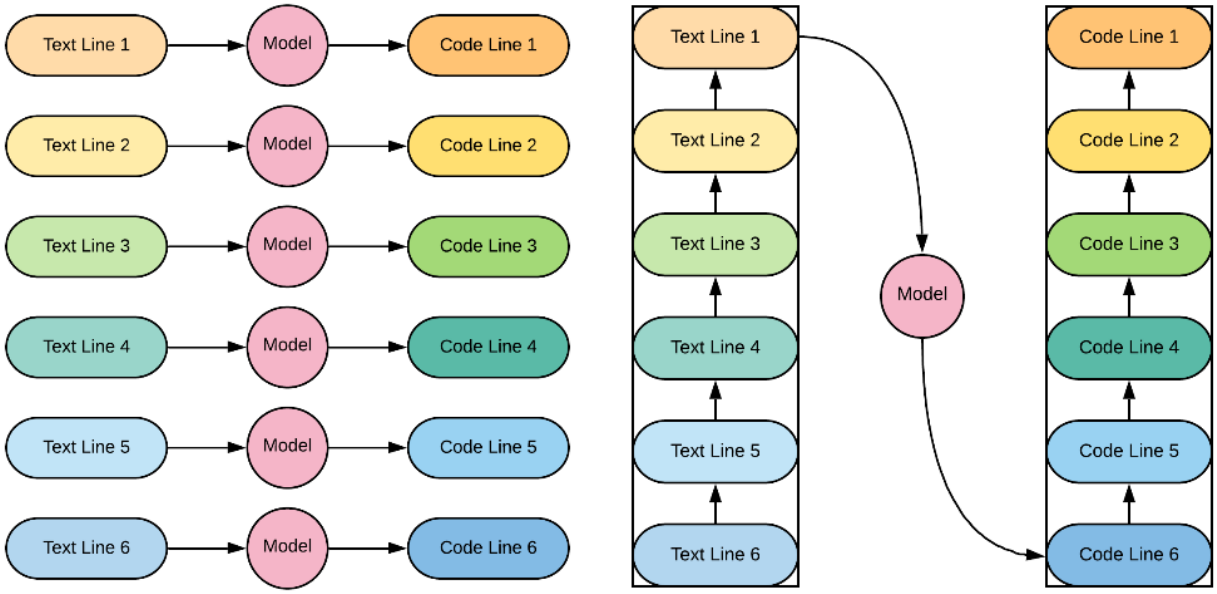
[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

[b]The ICD-10 code related to line 4 has been moved to line 3 by a human coder. Concatenating lines in a backward fashion restores alignment while preserving ordering.

[c]N/A: not applicable; the code that was previously here was moved to line 3, leaving this line blank.

In order to bypass this critical flaw in the investigated data set, a decision was taken to consider as input and target variables the certificate lines concatenated in a backward fashion (from part 2 to line 1 in part 1), as can be seen in Figure 1. This slight change in data format does not significantly alter the problem at hand, as the investigated model is still trained to recognize ICD-10–encoded medical entities from natural language. If anything, the modified modeling problem can be expected to be more difficult, as both the variance and dimensionality of both input and target variables have increased. Several methods are available to retrieve line-level aligned predictions from a model trained in such a configuration, for instance, using a combination of transfer learning and pruned tree search.

Figure 1. The original modeling problem and the modified investigated problem. In the original modeling problem (left), each certificate line is taken as an input variable to predict its corresponding ICD-10 code line. In the modified investigated problem (right), all certificate lines are concatenated and taken as an input variable to predict the corresponding concatenated ICD-10 code line. Lines 1-5 are from part 1 of the death certificate, and line 6 is part 2 of the certificate. ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.



Miscellaneous Variables

From gender to place of birth, a death certificate contains various additional types of information on its subject besides the chain of events leading to death. As some of these items are typically used by both expert systems and human coders to detect ICD-10 entities in the chain of events, they present an interest as explanatory variables for the investigated predictive model. After consultation with expert coders, the following items available on French death certificates were selected as additional exogenous variables:

- Gender (two-state categorical variable)
- Year of death (six-state categorical variable)
- Age, categorized into 5-year intervals, with the exception of subjects less than 1 year of age, who were divided into two classes depending on whether they were more than 28 days of age

- Origin of the death certificate (two-state categorical variable, from either the electronic- or paper-based death certification pipeline).

Strictly speaking, the subject's year of passing should only have a limited effect on the relationship between natural language and its contained medical entities. However, the WHO-defined coding rules, as well as their interpretations by human coders, evolve slightly over the years. As a consequence, the model should benefit, in terms of predictive performance, from being able to differentiate between different years.

Similarly, the impact of the certificate's origin on the model's predictive power is not entirely obvious at first sight. However, the data entry process for the paper-based certificates is handled by humans through speech recognition technology. In addition, the data entry clerks are asked to apply a small set of normalization rules to the natural language. Electronic death certificates, however, are received directly from the medical practitioner as is. As a consequence, distribution shifts are to be expected from the paper- to electronic-based chain of events, and including this information as an explanatory variable might be beneficial for the model's predictive power.

Model Definition

With both the explanatory and target variables well defined, the investigated modeling problem can be defined as follows:

The elements of equation 1 are defined as follows:

- $P(X)$ is the probability density of discrete random variable X
-  is the sequence of ICD-10 codes present on the death certificate concatenated on a single line of sequence length I
-  is the line in natural language, tokenized with a vocabulary V and of sequence length L
-  is the categorized age
-  is the year of death
-  is the gender
-  is the death certificate's origin
- f_θ is a mapping from the problem's input space to its output space, parameterized in $\theta \in R^n$, a real-valued vector (typically a neural network) of dimensionality $n \in \mathbb{N}$, the model's dimensionality.

Theoretically, the derived modeling problem is typical of traditional statistical modeling problems and could be solved using multinomial logistic regression. In practice, however, this approach presents a significant drawback. In this setting, the investigated target variable constitutes a categorical variable with 5616^{20} distinct states—death certificates in the data set have, at most, 20 ICD-10 codes in them, each of which can take 5616 distinct values—thus rendering the analysis intractable, both in terms of computational expanses and sample size requirements. This type of approach, however, makes no use of the data's inherent sequential nature, which allows the rewriting of the investigated modeling problem as follows:

where  is the i-th code present on the code line.

Factors in the right-hand side of equation 2 can be interpreted as constituting a distinct predictive modeling problem, all with an output variable distributed across all ICD-10 codes. Although still highly dimensional, predicting output variables of such dimensionality is typically tractable with modern machine learning techniques [7]. However, they present two significant drawbacks for traditional modeling techniques: (1) the number of output variables to predict varies across observations in the data set (not all death certificates have 20 ICD-10 codes) and (2) the output variables' distributions are conditioned on previous ones.

This particular formulation is known in the deep artificial neural network community as a Seq2Seq modeling problem [7] and has been an active area of research for the past few years. As one of the state-of-the-art neural architectures devised in the field, the Transformer [9] was chosen as the predictive model investigated in the following experiments. It was recently outperformed by the Evolved Transformer [17], a variation on the former. However, both approaches were investigated and yielded similar results. The Transformer architecture was retained due to its availability of official and maintained implementations, and the final results further displayed were obtained using an ensemble of seven such models. Each of the ensemble models' hyper-parameters and individual performances are available in Tables S1 and S2 of Multimedia Appendix 1, respectively.

Several specificities in the previously defined modeling problem required small adaptations to the Transformer architecture. However, the authors feel that their complexity falls outside the scope of this paper. The interested reader will, however, find a complete description of these modifications as well as a visualization in Figure S1 of Multimedia Appendix 1.

Finally, the authors are aware that many other approaches to sequential learning architectures are available, and have already been used, in order to address the problem investigated in this paper. The current state of the art on French death certificates, for instance, uses a multi-label classification approach. The authors chose not to investigate those methods for several reasons.

First, the task of extracting ICD-10 codes from natural language on death certificates is only a preliminary step in the production of a mortality statistics pipeline. The final task in this process is to derive the underlying cause of death, from these ICD-10 codes, following a set of rules defined by the WHO. The choice of the underlying cause of death from this set of rules heavily depends on the codes' order in the certificate. As a consequence, it is of paramount importance that the model be able to output these codes in the proper order, which is simply unachievable with a multiclass classification approach; this makes the problem a sequential learning problem, as our output is, indeed, a

sequence of variable-length tags taken from a set of well-defined classes. However, several approaches other than Seq2Seq are still available to solve such problems, such as connectionist temporal classification, which is typically used in optical character recognition tasks.

Second, the ICD-10 codes that the model needs to output are not necessarily independent. For instance, the presence of a given code in the outputted sequence can significantly alter other codes present in the sequence. As an example given by our expert coder, hematoma-related codes can be found in two ICD-10 chapters: first in chapter 9 of the ICD-10 classification (ie, codes related to circulatory diseases, beginning with an "I") and then in chapter 19 (ie, codes related to injury, poisoning, and certain other consequences of external causes, beginning with an "S" or a "T"). The choice of attributing the presence of the entity "hematoma" on a death certificate to the first or second possible chapter depends on whether an external cause—meaning an ICD-10 code from chapter 20—has already been outputted previously while converting the death certificate into codes. In order to account for such dependencies, we are compelled to model the joint distribution of the output sequence conditioned on the input variables, which is exactly what Seq2Seq is about. Therefore, the choice of using Seq2Seq approaches to solve the modeling problem investigated in this paper becomes not only natural but almost compulsory. In addition, due to the data-driven tokenization used in order to make use of the ICD-10 classification's hierarchical nature, some tokens that the model is allowed to predict are not valid ICD-10 codes. For instance, the code "I659" could be decomposed into a sequence of two codes (ie, "I65" and "9-" with the "-" character at the end used to keep track of spaces between codes). It appears clear here that when the model needs to output an "I659" code, predicting "9-" in itself is not possible without any conditioning on "I65" appearing earlier.

Training and Evaluation Methodology

The investigated model was trained using all French death certificates from the year 2011 to 2016. A total of 5000 certificates were randomly excluded from each year; these were distributed into a validation set for hyper-parameter fine-tuning and into a test data set for unbiased prediction performance estimation (2500 certificates each), resulting in three data sets with following sample sizes: (1) training data set (3,240,109 records), (2) validation data set (30,000 records), and (3) test data set (30,000 records).

The model was adapted from TensorFlow's official Transformer implementation; TensorFlow is a Python-based distributed machine learning framework. Training was performed on three NVIDIA RTX 2070 GPUs simultaneously with a mirrored distribution strategy using a variant of stochastic gradient descent, the Adam optimization algorithm.

Hyper-parameters were first initialized following the Transformer's base setting, according to the architecture's authors. Further fine-tuning of a selected number of hyper-parameters was performed using a random search guided on the validation set. The interested reader will find a complete description of the training process and hyper-parameter values defining this model in Multimedia Appendix 1.

After training, the model's predictive performance was assessed on the test data set, which was excluded prior to training, as mentioned earlier, and compared to the current state of the art, obtained by the LIMSI during the 2017 CLEF eHealth challenge [2]. As the CLEF eHealth challenge only provided electronic certificates to the contestants, and in order to ensure comparability, the model's performance was assessed using paper-based and electronic certificates, separately. For the same reason, the performance metrics used for model evaluation were selected as follows:

The elements of equations 4 and 5 are defined as follows:

- True positives: the number of codes predicted by the model that are present in the test set's true output target
- False positives: the number of codes predicted by the model that are not present in the test set's true output target
- False negatives: the number of codes not predicted by the model that are present in the test set's true output.

Note that predictions are considered as true positives only for exact code matches, up to the fourth character. Table 3 shows an example of how this can affect the reported performance, by focusing on a line of the causal chain of events leading to death reported in Table 1 and fictional examples of predictions, as follows:

- The first prediction example outputs two incorrect codes. The number of true positives is, thus, 0, leading to all metrics being evaluated as 0.
- The second prediction example correctly outputs the first code (I64: "Stroke") but fails to correctly output the second code's fourth character (G81: "Hemiplegia" is predicted instead of the ground-truth value G819: "Hemiplegia, unspecified"). Although the prediction and ground truth are quite similar (ie, they share the three first characters), this code is considered incorrect, which leads to counts of both one false positive (ie, the code was predicted incorrectly) and one false negative (ie, the correct G819 code was not predicted), leading to all metrics being evaluated as 0.5.
- The third prediction example correctly outputs the first code but fails to recognize any additional codes from the textual input, leading to a precision of 1 (ie, all predicted codes are indeed true positives) and a recall of 0.5 (ie, one code present in the ground truth was not predicted). This then leads to an F-measure of 0.66. Note that in this context, the F-measure is higher than in the second example.
- The fourth prediction example correctly outputs both codes but also outputs two additional and completely unrelated codes, leading to a precision of 0.5 (ie, only half of the predicted codes are present in the ground truth) and a recall of 1 (ie, all codes present in the ground truth were correctly predicted), leading to an F-measure of 0.66.
- The fifth prediction example correctly outputs both codes and does not predict any additional codes (ie, perfect prediction), leading to all metrics being evaluated as 1.
- The sixth prediction example correctly outputs both codes and does not predict any additional codes. However, the

codes are in the wrong order, but this is not penalized in any way in the metrics definitions, so this prediction is associated with metrics all being evaluated as 1.

Table 3. Examples of how the selected performance metrics behave for different predictions. The input text was "stroke in September left hemiparesis" and the true ICD-10 encoding was I64 and G819.

Prediction example	ICD-10[a] codes	Precision	Recall	F-measure
1	B189 H155	0.0	0.0	0.0
2	I64 G81	0.5	0.5	0.5
3	I64	1.0	0.5	0.66
4	I64 G819 A338 B87	0.5	1.0	0.66
5	I64 G819	1.0	1.0	1.0
6	G819 I64	1.0	1.0	1.0

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

The informed reader might find that these metrics stray away from common machine translation system benchmarking metrics, such as bilingual evaluation understudy (BLEU) or negative log perplexity scores [7-9,18], but the former were the only ones used in comparable work. As BLEU and negative log perplexity have close to no absolute interpretability without comparisons to alternative methods, their use was discarded from the experiment. In order to present the reader with a more comprehensive view of the performance of the proposed approaches, these accuracy metrics were also derived on a per-chapter basis, again on the same test set, and 95% CIs were computed using bootstrapping.

Results

Performance Evaluation

The ensemble of Transformer models were trained as previously described for approximately 3 weeks; the final ensemble's predictive performance and that of the current state-of-the-art model are reported in Table 4. As previously mentioned, the performance of the current state-of-the-art model was assessed based on electronic certificates only and should, as a consequence, be compared to the performance of the proposed approach based on a similar situation. Because paper-based certificates are still more common than their electronic counterparts in France (ie, approximately 90% of certificates in the data set are paper based), the performance of the approach using all certificates and that of the paper-based certificate approach are also displayed.

Table 4. Assessments of the current state-of-the-art model and the proposed approach.

Approach	F-measure (95% CI)[a]	Precision (95% CI)	Recall (95% CI)
Current state of the art: LIMSI[b]	0.825[c]	0.872[c]	0.784[c]
Proposed approach: electronic certificates	0.952 (0.946-0.957)	0.955 (0.95-0.96)	0.948 (0.943-0.954)
Proposed approach: paper-based certificates	0.942 (0.941-0.944)	0.949 (0.947-0.95)	0.936 (0.934-0.937)
Proposed approach: all certificates	0.943 (0.941-0.944)	0.949 (0.948-0.951)	0.937 (0.935-0.938)

[a]95% CIs were derived by bootstrapping.

[b]LIMSI: Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur.

[c]95% CIs were not provided in the LIMSI's publication and are, therefore, not displayed.

The proposed approach shows an F-measure that is 73% closer to a perfect score when compared to the current state-of-the-art approach. In addition to its substantial improvement in the F-measure, the proposed approach displays significantly more balanced precision and recall scores than the LIMSI's method: from 5% relative difference to less than 1%.

A surprising result, however, lies in the model's lower performance based on paper-based certificates. Indeed, the standardization they receive due to their voice-based data collection process considerably reduces variance and prevents any misspelled words in the data that are potentially present in electronic-based certificates. As a consequence, model performance on the former should be expected to be higher. A potential explanation for this phenomenon lies in the potential for missing data in paper-based certificates. Indeed, when confronted with poorly written words, data clerks are allowed to replace them with the "!" symbol when the word is estimated to be unreadable; this occurs in approximately 10% of paper-based certificates. Medical coders, however, are usually more efficient in guessing the words from the written certificates, typically with the addition of contextual clues. A purely text-based approach, however, is then limited to pure guesses for those observations with missing data, logically leading to poorer performance. Because this phenomenon is absent from electronic-based certificates, it is a promising candidate for explaining this unexpected difference in

performance. In addition, the model performance based on paper-based certificates that did not contain any "!" symbols in the test set led to an F-measure of 96.2%, thus providing strong evidence to support this hypothesis.

Per-Chapter Quantitative Analysis

Although the proposed approach significantly outperformed the current state-of-the-art approach, neural network–based methods are known to present several drawbacks that can significantly limit their application in some situations. Typically, the current lack of systematic methods to interpret and understand neural network–based models and their decision processes can lead the former to perform catastrophically on incorrectly predicted cases, independent from their high predictive performance. As a consequence, the proposed model behavior in incorrectly predicted cases requires careful analysis. In addition, such an investigation can lead to significant insights that are potentially relevant when applying the derived model in practical applications.

One simple, straightforward approach to understanding the model's weakness lies in assessing its performance on a finer-grain level, for instance, by identifying false positives and negatives not only at the global level, but per ICD-10 chapters, as can be seen in Table 5.

It appears from this table that although the most prevalent medical entities are associated with low false positive and negative rates, some rarer chapters are associated with unreasonably high error rates. Depending on their prevalence and accuracies, these chapters can be classified into two distinct categories:

1. Chapters associated with unreasonably high error rates but extremely low prevalence, such as "Diseases for the ear and mastoid process" or "Pregnancy, childbirth and the puerperium." However, these entity groups remain rare enough within the data set to allow for alternative treatments, like manual evaluation, for instance.
2. Chapters associated with high error rates, although lower than the former, but with significant prevalence, such as "External causes of morbidity and mortality" or "Injury, poisoning and certain other consequences of external causes."

The task of identifying these potential mistakes, however, is not entirely trivial depending on whether mistakes are of false positive or false negative types. Indeed, potential false positive errors are directly identifiable within the predicted ICD-10 code sequences. As a consequence, coding quality control for this type of mistake should be fairly straightforward to implement: one could, for instance, manually review all code sequences containing codes related to "Pregnancy, childbirth and the puerperium" systematically. Potential false negative errors, however, are inherently significantly harder to identify and require further investigation, for instance, through association rules analysis.

A number of promising leads are already available and should reasonably improve upon the proposed approach:

- Training methods adapted to imbalanced data sets, such as up-sampling or loss weighting
- Data augmentation for rare medical entities
- Addition of information to the model (ie, prenatal-related death, for instance, is explicitly defined as such on certificates)
- A hybrid approach with traditional NLP approaches, which are typically less expensive in terms of sample size requirements.

Table 5. False positive, false negative, and prevalence rates for each ICD-10 chapter, sorted in descending order by prevalence.

ICD-10[a] chapter	False positives, %	False negatives, %	Prevalence, %
Diseases of the circulatory system	3.75	4.98	22.4
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	3.87	4.12	21.8
Neoplasms	4.07	5.07	15.9
Diseases of the respiratory system	3.02	4.00	8.76
Endocrine, nutritional and metabolic diseases	2.17	3.44	4.83
Diseases of the nervous system	2.70	4.12	3.89
Mental, behavioral and neurodevelopmental disorders	2.88	4.14	3.58
Diseases of the digestive system	5.72	8.10	3.53
Factors influencing health status and contact with health services	19.2	19.6	3.08
Diseases of the genitourinary system	5.45	7.59	2.71
External causes of morbidity and mortality	16.6	23.5	2.57
Certain infectious and parasitic diseases	7.98	9.23	2.55
Injury, poisoning and certain other consequences of external causes	14.0	19.8	2.07
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	6.72	12.2	0.77
Diseases of the musculoskeletal system and connective tissue	12.2	17.3	0.62
Diseases of the skin and subcutaneous tissue	8.72	8.16	0.51
Certain conditions originating in the perinatal period	14.5	20.5	0.16
Congenital malformations, deformations and chromosomal abnormalities	22.4	25.6	0.15
Diseases of the eye and adnexa	4.93	13.6	0.076
Codes for special purposes	24.0	34.0	0.047
Diseases of the ear and mastoid process	5.60	33.3	0.017
Pregnancy, childbirth and the puerperium	50.0	33.3	0.0056

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

Score Calibration Fitness Assessment

When the model is fit in a similar fashion to multinomial logistic regression, it not only yields a prediction but an associated score similar to a confidence probability. If properly calibrated, this score can offer powerful insights regarding the prediction's quality at the individual level. Typically, a "good" score would be expected to show higher values in cases where the ICD-10 sequence is correctly predicted and lower values when incorrectly predicted. Such a well-calibrated score could, for instance, allow for real-world applications of semiautonomous systems where the following occurs:

- A threshold value for the model's score is defined.
- All certificates whose predictions are associated with confidence scores above the threshold level are accepted without any additional human supervision.
- All certificates whose predictions are associated with confidence scores below the threshold level are systematically reviewed by a human expert and modified manually, if required.

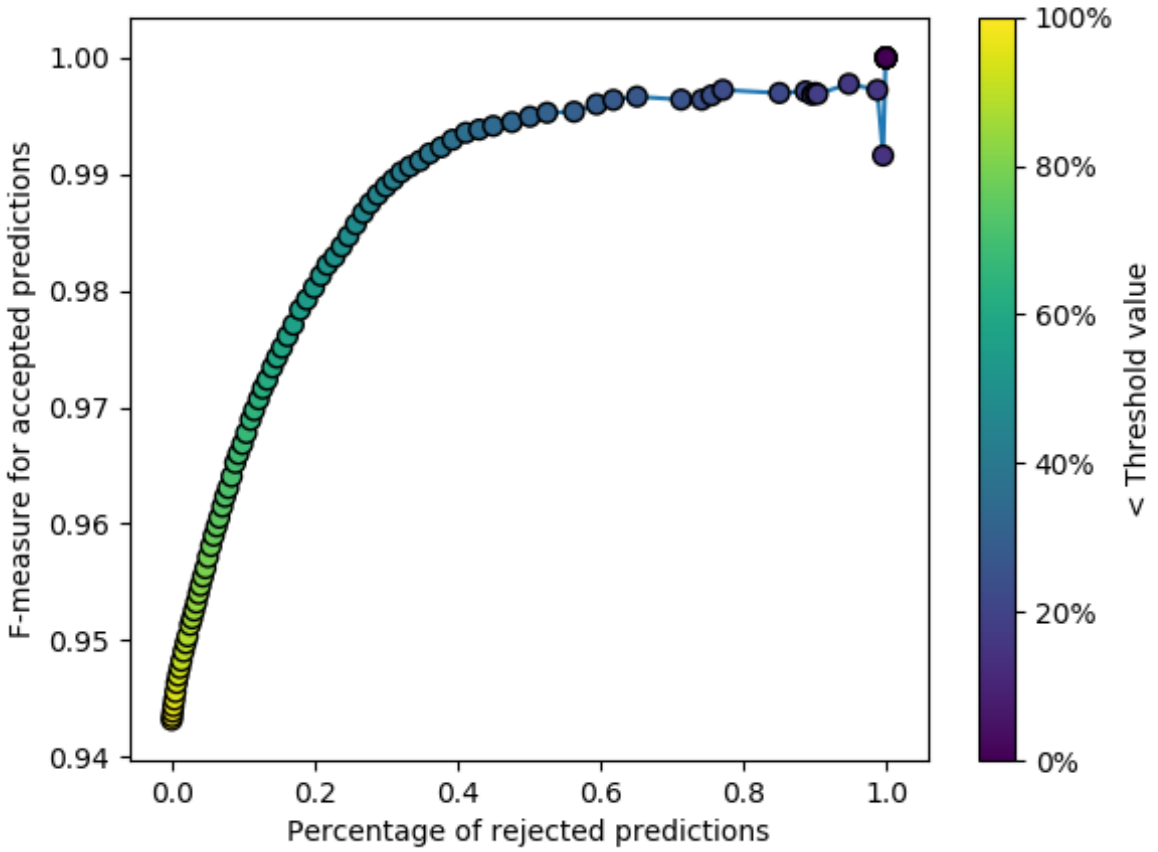
Being able to properly filter the model's predictions according to a well-calibrated confidence score would, thus, allow us to get the best of both worlds. Most of the certificates would be automatically coded by the autonomous system, leaving human coders with only the most complex cases.

Efficient assessment of such scores in traditional machine learning problems is typically done through visualization of receiver operating characteristic (ROC) curves. However, the sequential multinomial nature of the investigated problem renders this approach ill-defined. The plot in Figure 2, while conceptually similar to an ROC curve, was derived following a slightly different approach in order to efficiently appreciate the model score's quality. This visualization was derived as follows:

- A grid of score threshold values was defined with a uniform grid with 0.01 intervals, corresponding to the threshold

defined above, filtering between model predictions that would require human examination or not.
- For every given threshold value, we computed the percentage of predictions with inferior or equal scores, which were considered as rejected, requiring human examination due to poor score; we also computed the F-measure performance on the predictions with high enough scores that would be accepted without any human intervention following the above example.
- The percentage of accepted certificates and F-measures were plotted on a scatterplot against each other, with threshold values displayed as colored points.

Figure 2. Percentage of rejected predictions versus F-measure for accepted ones. The score threshold values defining the accepted predictions are displayed as colored points.



By showing a clean, increasing relationship between the number of rejected predictions and the F-measure evaluated using the remaining certificates, Figure 2 strongly indicates good score calibration. As an example, by considering that only predictions associated with a confidence score lower than 0.5 do not require any additional human supervision, the system is able to code approximately 80% of all certificates present in the test set with an F-measure of 0.98, significantly higher than the value of 0.94 obtained on all test certificates.

Discussion

Principal Findings

The error analysis carried out so far allowed for the assessment of the model's strengths and weaknesses on a global level. However, it failed to yield any interesting insights regarding potential model biases, for instance, toward specific coding rules. Indeed, the coding of medical entities from natural language, especially with regard to mortality statistics, is subject to a number of coding rules depending on context or pathology, with a level of specificity oftentimes reaching casuistry [1].

In addition, all results have been presented so far with a model error defined as a disagreement between the model's output and the information contained in the database. However, building a medical database is a complex, mostly human-based process. As such, an inevitable amount of noise is to be expected in the ICD-10 codes present in the database, in two main forms. The first form is simple human errors in the ICD-10. The second form is the presence of unreadable text in paper-based certificates. Unreadable words on paper-based certificates are denoted as an exclamation point in the textual data that is fed to the model. However, human coders usually take additional time to infer these words, for instance, via queries to the medical certifier or from contextual cues. This leads to death certificates in the database where the ICD-10 sequences contain additional codes compared to the textual data available. As such, not predicting these codes would result in a drop in performance metrics, while the model has no way of predicting them. An

example of such a death certificate can be found in Table S3 in Multimedia Appendix 1. These phenomena have the potential to negatively bias the proposed model's performance estimations and should be the object of further investigation.

One straightforward, although fairly time-consuming, approach to address these two considerations can be to have an ICD-10 coding expert manually examine some of the death certificates where the model's predictions do not match the ICD-10 codes present in the database. Two experiments were conducted following this idea.

In the first experiment, 99 certificates where the model's predictions did not exactly match with the database's ICD-10 variables (ie, the ICD-10 sequences differed by at least one code) were selected at random from the test set and were shown to the medical practitioner representative and final decision maker on ICD-10 mortality coding in France, who was asked to do the following with each certificate:

- Manually recode all the ICD-10 medical entities present on each death certificate by herself using the information the proposed model had access to, without access to the data set or to the model's proposed ICD-10 sequences.
- Give a qualitative comment on the outputs of the investigated model and database as compared to hers.

Since the ICD-10 sequences derived from the medical expert and national representative for ICD-10 coding in France are significantly more reliable than the ones coming from the traditional data production process (ie, using a combination of expert system and human coders), they can be considered as exempt of any potential human error. As a consequence, comparing them to both the proposed model's output and the ICD-10 values contained in the data set would allow for an estimation of the potential negative biases described above. This can be done, for instance, by estimating the performance metrics selected for the previous experiments, considering both the model's predictions and the database's values as predictions, and the medical expert's outputs as the ground truth. Depending on the resulting values, several interpretations can be made ranging between two extreme cases:

1. If perfect agreement (ie, an F-measure of 1.0) is reached between the database's ICD-10 sequences and the medical expert's outputs, suggesting that the database does not have any coding mistakes, then the performance metrics reported in the Results section can safely be considered unbiased.
2. If perfect agreement is not reached between the model's predictions of ICD-10 sequences and the medical expert's outputs, suggesting that the model did not make any mistakes, then the performance metrics reported in the Results section should be considered significantly underestimated.

However, before estimating the performance metrics following this methodology, a slight preprocessing step is required. Indeed, on the death certificates sampled for the experiment, the F-measure estimation between the model's prediction and the database's ICD-10 sequences yielded a value of 0.81. This is explained by the sampling process, in which death certificates were selected where at least one code differed in both ICD-10 sequences. As a consequence, and because of the model's performance, most ICD-10 codes present on both sequences were identical, as can be seen with the error examples presented in Tables S3 to S5 in Multimedia Appendix 1. The authors felt that this might lead to artificially high values of the estimated metrics in the experiment; consequently, we decided to delete all common codes on both the model's outputs and the database's values prior to metrics estimation, as shown in Table 6.

For better comparability, these statistics are reported based on both (1) certificates without missing data in the natural language–based causal chain of events leading to death (by excluding certificates containing the "!" symbol) in Table 7 and (2) all certificates in Table 8.

Table 6. Example of preprocessing used for the experiment on a real error example. The predicted and database ICD-10 sequences only differ by one code, while they share five codes. All shared codes were deleted from all ICD-10 sequences prior to estimation of performance metrics.

Source of ICD-10[a] codes	ICD-10 codes before preprocessing	ICD-10 codes after preprocessing
Predicted by the model	I259 Z951 I719 C679 I10 R092	Z951
Present in the database	I259 I251 I719 C679 I10 R092	I251
Predicted by medical expert	I259 I251 I719 C679 I10 R092	I251

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

Table 7. F-measure, precision, and recall of both the database ICD-10 codes and the model's prediction of codes compared to that of the medical expert for sampled certificates without missing data.

Source of ICD-10[a] codes	F-measure (95% CI)	Precision (95% CI)	Recall (95% CI)
Presence in database against medical expert prediction	0.483 (0.383-0.589)	0.443 (0.341-0.555)	0.531 (0.425-0.636)
Model prediction against medical expert prediction	0.431 (0.316-0.542)	0.458 (0.338-0.580)	0.407 (0.295-0.519)

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

Table 8. F-measure, precision, and recall of both the database ICD-10 codes and the model's prediction of codes compared to that of the medical expert for all sampled certificates.

Source of ICD-10[a] codes	F-measure (95% CI)	Precision (95% CI)	Recall (95% CI)
Presence in database against medical expert prediction	0.613 (0.486-0.733)	0.630 (0.492-0.761)	0.596 (0.471-0.721)
Model prediction against medical expert prediction	0.370 (0.237-0.504)	0.392 (0.250-0.540)	0.351 (0.222-0.482)

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

Tables 7 and 8 show no significant difference in prediction performance between the proposed approach and the current data production process (ie, based on a combination of expert system and human coders), although the database's ICD-10 values have better performance metrics in both cases. When including certificates containing missing text, the proposed model's agreement with the medical expert increases considerably, further confirming the hypothesis that the performance metrics reported in the Results section were negatively biased.

From the qualitative comments made by the medical expert, three major types of model errors could be defined:

1. In 16% (16/99) of cases, disagreement between the current data production process and the proposed approach was due to missing information in the input text. On these specific cases, the F-measure between the model's output and medical expert's decision was determined to be 0.974; an example of such an error case can be seen in Table S3 in Multimedia Appendix 1.
2. In 14% (14/99) of cases, the correct ICD-10 sequence was dependent on highly contextual clues or external knowledge of world behavior (eg, someone found dead at the bottom of a set of stairs is quite likely to have suffered a fall). An example of such an error case can be seen in Table S4 in Multimedia Appendix 1.
3. In 12% (12/99) of cases, the correct ICD-10 sequence was dependent on highly nonlinear, almost casuistic rules. These were typical examples of scenarios where a hybridized deep learning and expert-based system would be beneficial; an example of such an error case can be seen in Table S5 in Multimedia Appendix 1.

The remaining cases did not elicit any comment from the medical expert.

Finally, in the second experiment, the medical expert's ability to discriminate between human coding and the proposed approach was assessed in a Turing test-like approach. To do so, 100 additional certificates where the model's output differed from the database's ICD-10 sequences were sampled at random from the test set. The medical expert was shown their corresponding input features (ie, text and auxiliary variables) as well as the two ICD-10 sequences, with their provenance from either the model or the database masked, as can be seen in Table 9.

Table 9. Example of death certificate format given to the medical expert for the second experiment. The medical expert was asked, based on the information available in the line, to guess which of propositions 1 or 2 was produced by a human coder, with the other being the proposed model's output.

Item	Sex[a] of deceased	Year of death	Age of deceased (years)	Certificate text[b]	Proposition 1 (ICD-10[c] codes)	Proposition 2 (ICD-10 codes)
Death certificate	2	2013	90	90 ans, péritonite, perforation grêle, occlusion, chirurgie digestive, infection pulmonaire, arrêt respiratoire	R54 K566 K659, K631 Y839 J958 R092	R54 K659 K631 K566 Y839 J189 R092

[a]Sex is a two-state categorical variable: 1 (female) or 2 (male).

[b]The certificate text was taken from a death certificate in France and is, therefore, written in French.

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision.

After exclusion of certificates containing missing text data, where the human coder was easily identifiable due to the apparently out-of-context additional codes (Table S3 in Multimedia Appendix 1), the medical expert was able to correctly identify the human coder in 63% (62/99; 95% CI 50.7%-73.2%) of cases, which is significantly better than random guessing, although barely.

Conclusions

In this paper, the task of automatic recognition of ICD-10 medical entities from natural language in French was presented as a Seq2Seq modeling problem, well known in the deep artificial neural network academic literature. From this consideration, the performance of a well-known approach in the field, consisting of an ensemble of Transformer models, was investigated using the CépiDc database and was shown to reach a new state of the art. The derived model's behavior was thoroughly assessed following different approaches in order to identify potential weaknesses and elements for improvements. Although the proposed approach significantly outperformed any other existing automated ICD-10 recognition systems based on French free text, the question of method transferability to other languages requires more investigation.

The substantial performance reported in this paper makes possible a range of promising applications in various

medical-related fields, from automated medical coding to advanced natural language–based analysis for epidemiology. However, these interesting opportunities are oftentimes prohibited by these methods' massive drawbacks, mostly their requirement for millions of annotated observations in order to perform well. Mortality data sets, despite their specificity, provide researchers with a huge amount of clean, multilingual medical text data perfectly fit for the application of deep neural networks. As a consequence, and keeping in mind the strong transfer learning capability of neural networks, the authors firmly believe that mortality data constitute one of the most promising points of entry into modern NLP methods applications in the biomedical sciences.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File , 129 KB - medinform_v10i4e26353_app1.docx]

References

1. International Statistical Classification of Diseases and Related Health Problems, 10th Revision. Geneva, Switzerland: World Health Organization; 2019. URL: https://icd.who.int/browse10/2019/en#/ [accessed 2022-03-19]
2. Névéol A, Robert A, Anderson R, Cohen B, Grouin C, Lavergne T, et al. CLEF eHealth 2017 Multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: Proceedings of the Workshop of the Cross-Language Evaluation Forum, CEUR Workshop Proceedings. 2017 Presented at: Workshop of the Cross-Language Evaluation Forum, CEUR Workshop; January 2017; Dublin, Ireland p. 1-17 URL: http://ceur-ws.org/Vol-1866/invited_paper_6.pdf
3. Névéol A, Bretonnel Cohen K, Grouin C, Hamon T, Lavergne T, Kelly L, et al. Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Proceedings of the 7th International Conference of the CLEF Association, CEUR Workshop Proceedings. 2016 Presented at: The 7th International Conference of the CLEF Association, CEUR Workshop; September 5-8, 2016; Évora, Portugal p. 28-42 URL: http://ceur-ws.org/Vol-1609/16090028.pdf
4. Duarte F, Martins B, Sousa Pinto C, Silva MJ. A deep learning method for ICD-10 coding of free-text death certificates. In: Proceedings of the 18th EPIA Conference on Artificial Intelligence. 2017 Presented at: The 18th EPIA Conference on Artificial Intelligence; September 5-8, 2017; Porto, Portugal p. 137-149. [doi: 10.1007/978-3-319-65340-2_12]
5. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: The 26th International Conference on Neural Information Processing Systems; December 5-10, 2013; Lake Tahoe, NV p. 3111-3119 URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
6. Devin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language onderstanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf [doi: 10.18653/v1/N19-1423]
7. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: The 2014 Conference on Empirical Methods in Natural Language Processing; October 25-29, 2014; Doha, Qatar p. 1724-1734 URL: https://aclanthology.org/D14-1179.pdf [doi: 10.3115/v1/d14-1179]
8. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convoutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning. 2017 Presented at: The 34th International Conference on Machine Learning; August 6-11, 2017; Sydney, Australia p. 1243-1252 URL: http://proceedings.mlr.press/v70/gehring17a/gehring17a.pdf
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 5998-6008 URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
10. Ševa J, Sänger M, Leser U. WBI at CLEF eHealth 2018 Task 1: Language-independent ICD-10 coding using multi-lingual embeddings and recurrent neural networks. In: Proceedings of 9th International Conference of the CLEF Association, CEUR Workshop Proceedings. 2018 Presented at: The 9th International Conference of the CLEF Association, CEUR Workshop; September 10-14, 2018; Avignon, France p. 1-14 URL: http://ceur-ws.org/Vol-2125/paper_118.pdf

11. Atutxa A, de Ilarraza AD, Gojenola K, Oronoz M, Perez-de-Viñaspre O. Interpretable deep learning to map diagnostic texts to ICD-10 codes. Int J Med Inform 2019 Sep;129:49-59. [doi: 10.1016/j.ijmedinf.2019.05.015] [Medline: 31445289]
12. Reys AD, Sylva D, Severo D, Pedro S, de Sousa MM, Salgado GAC. Predicting multiple ICD-10 codes from Brazilian-Portuguese clinical notes. In: Proceedings of the Brazilian Conference on Intelligent Systems. 2020 Presented at: The Brazilian Conference on Intelligent Systems; October 20-23, 2020; Rio Grande, Brazil p. 566-580. [doi: 10.1007/978-3-030-61377-8_39]
13. Cao P, Yan C, Fu X, Chen Y, Liu K, Zhao J, et al. Clinical-coder: Assigning interpretable ICD-10 codes to Chinese clinical notes. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020 Presented at: The 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; July 5-10, 2020; Virtual p. 294-301 URL: https://aclanthology.org/2020.acl-demos.33.pdf [doi: 10.18653/v1/2020.acl-demos.33]
14. Schafer H, Friedrich CM. Multilingual ICD-10 code assignment with transformer architectures using MIMIC-III discharge summaries. In: Proceedings of the 11th Conference and Labs of the Evaluation Forum. 2020 Presented at: The 11th Conference and Labs of the Evaluation Forum; September 22-25, 2020; Virtual p. 1-16 URL: http://ceur-ws.org/Vol-2696/paper_212.pdf
15. Amin S, Neumann G, Dunfield K, Vechkaeva A, Chapman KA, Wixted MK. MLT-DFKI at CLEF eHealth 2019: Multi-label classification of ICD-10 codes with BERT. In: Proceedings of the 10th Conference and Labs of the Evaluation Forum. 2019 Presented at: The 10th Conference and Labs of the Evaluation Forum; September 9-12, 2019; Lugano, Switzerland p. 1-15 URL: http://ceur-ws.org/Vol-2380/paper_67.pdf
16. HM Passport Office. Completing a medical certificate of cause of death (MCCD). GOV.UK. 2018 Sep 25. URL: https://www.gov.uk/government/publications/guidance-notes-for-completing-a-medical-certificate-of-cause-of-death [accessed 2019-04-24]
17. So D, Le Q, Liang C. The evolved transformer. In: Proceedings of the 36th International Conference on Machine Learning. 2019 Presented at: The 36th International Conference on Machine Learning; June 9-15, 2019; Long Beach, CA p. 5877-5886 URL: http://proceedings.mlr.press/v97/so19a/so19a.pdf
18. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Presented at: The 40th Annual Meeting on Association for Computational Linguistics; July 7-12, 2002; Philadelphia, PA p. 311-318 URL: https://dl.acm.org/doi/pdf/10.3115/1073083.1073135 [doi: 10.3115/1073083.1073135]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BLEU: bilingual evaluation understudy
CépiDc: Centre for Epidemiology on Medical Causes of Death
CLEF: Conference and Labs of the Evaluation Forum
ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision
LIMSI: Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
NLP: natural language processing
ROC: receiver operating characteristic
Seq2Seq: sequence-to-sequence
SVM: support vector machine
WHO: World Health Organization

Edited by C Lovis; submitted 08.12.20; peer-reviewed by V Montmirail, H Park; comments to author 10.01.21; revised version received 23.02.21; accepted 08.01.22; published 11.04.22.

Please cite as:
Falissard L, Morgand C, Ghosn W, Imbaud C, Bounebache K, Rey G
Neural Translation and Automated Recognition of ICD-10 Medical Entities From Natural Language: Model Development and Performance Assessment
JMIR Med Inform 2022;10(4):e26353
URL: https://medinform.jmir.org/2022/4/e26353
doi:10.2196/26353
PMID:35404262

Original Paper

Natural Language Processing for Assessing Quality Indicators in Free-Text Colonoscopy and Pathology Reports: Development and Usability Study

Jung Ho Bae[1,2,3], MD; Hyun Wook Han[1,2*], MD, PhD; Sun Young Yang[3*], MD, PhD; Gyuseon Song[1,2], BS; Soonok Sa[1,2], MS; Goh Eun Chung[3], MD, PhD; Ji Yeon Seo[3], MD, PhD; Eun Hyo Jin[3], MD, PhD; Heecheon Kim[4], MS; DongUk An[4], PhD

[1]Department of Biomedical Informatics, CHA University School of Medicine, CHA University, Seongnam, Republic of Korea

[2]Institute for Biomedical Informatics, CHA University School of Medicine, CHA University, Seongnam, Republic of Korea

[3]Department of Internal Medicine and Healthcare Research Institute, Healthcare System Gangnam Center, Seoul National University Hospital, Seoul, Republic of Korea

[4]Miso Info Tech Co, Ltd, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hyun Wook Han, MD, PhD
Department of Biomedical Informatics
CHA University School of Medicine
CHA University
335, Pangyo-ro, Bundang-gu
Seongnam, 13488
Republic of Korea
Phone: 82 31 881 7109
Fax: 82 31 881 7069
Email: stepano7@gmail.com

Abstract

Background: Manual data extraction of colonoscopy quality indicators is time and labor intensive. Natural language processing (NLP), a computer-based linguistics technique, can automate the extraction of important clinical information, such as adverse events, from unstructured free-text reports. NLP information extraction can facilitate the optimization of clinical work by helping to improve quality control and patient management.

Objective: We developed an NLP pipeline to analyze free-text colonoscopy and pathology reports and evaluated its ability to automatically assess adenoma detection rate (ADR), sessile serrated lesion detection rate (SDR), and postcolonoscopy surveillance intervals.

Methods: The NLP tool for extracting colonoscopy quality indicators was developed using a data set of 2000 screening colonoscopy reports from a single health care system, with an associated 1425 pathology reports. The NLP system was then tested on a data set of 1000 colonoscopy reports and its performance was compared with that of 5 human annotators. Additionally, data from 54,562 colonoscopies performed between 2010 and 2019 were analyzed using the NLP pipeline.

Results: The NLP pipeline achieved an overall accuracy of 0.99-1.00 for identifying polyp subtypes, 0.99-1.00 for identifying the anatomical location of polyps, and 0.98 for counting the number of neoplastic polyps. The NLP pipeline achieved performance similar to clinical experts for assessing ADR, SDR, and surveillance intervals. NLP analysis of a 10-year colonoscopy data set identified great individual variance in colonoscopy quality indicators among 25 endoscopists.

Conclusions: The NLP pipeline could accurately extract information from colonoscopy and pathology reports and demonstrated clinical efficacy for assessing ADR, SDR, and surveillance intervals in these reports. Implementation of the system enabled automated analysis and feedback on quality indicators, which could motivate endoscopists to improve the quality of their performance and improve clinical decision-making in colorectal cancer screening programs.

(JMIR Med Inform 2022;10(4):e35257) doi:10.2196/35257

KEYWORDS

natural language processing; colonoscopy; adenoma; endoscopy

Introduction

High-quality colonoscopy is a proven method of reducing colorectal cancer risk by allowing early detection and removal of premalignant polyps [1]. However, there are considerable variations in the quality of colonoscopies performed by endoscopists [2-4]. Therefore, quality assurance is an essential part of colonoscopy screening programs, and the American Society of Gastrointestinal Endoscopy/American College of Gastroenterology Task Force on Quality in Endoscopy has published indicators for colonoscopy to improve safety and quality [5]. While all the indicators are important, the adenoma detection rate (ADR) and sessile serrated lesion (SSL) detection rate (SDR) of endoscopists are well-established key indicators of postcolonoscopy colorectal cancer incidence and related deaths [5-7]. Another crucial quality indicator is the adherence to guidelines for setting the frequency of follow-up colonoscopies, known as the surveillance interval. Recommending an incorrect surveillance interval may increase the incidence of metachronous lesion or lead to the overuse of colonoscopies [8].

Periodically reporting to endoscopists their performance on quality measures effectively improves the quality of colonoscopies by encouraging introspection and motivation for behavior changes [9-11]. However, reporting ADR, SDR, and surveillance intervals requires careful manual review of colonoscopy reports and their associated pathology reports and following this review with a calculation of polyp data based on clinical guidelines. This series of processes for quality reporting is laborious and time-consuming.

Natural language processing (NLP) is a computer-based linguistics technique used to extract information from free-text data documents [12]. NLP allows the automation of report creation by extracting important clinical information from unstructured free-text documents. NLP has been used in various clinical fields [12-17]. The application of NLP to information extraction requires identifying clinical information, such as adverse events, and facilitates various aspects of optimizing clinical work, such as quality control and patient management [18].

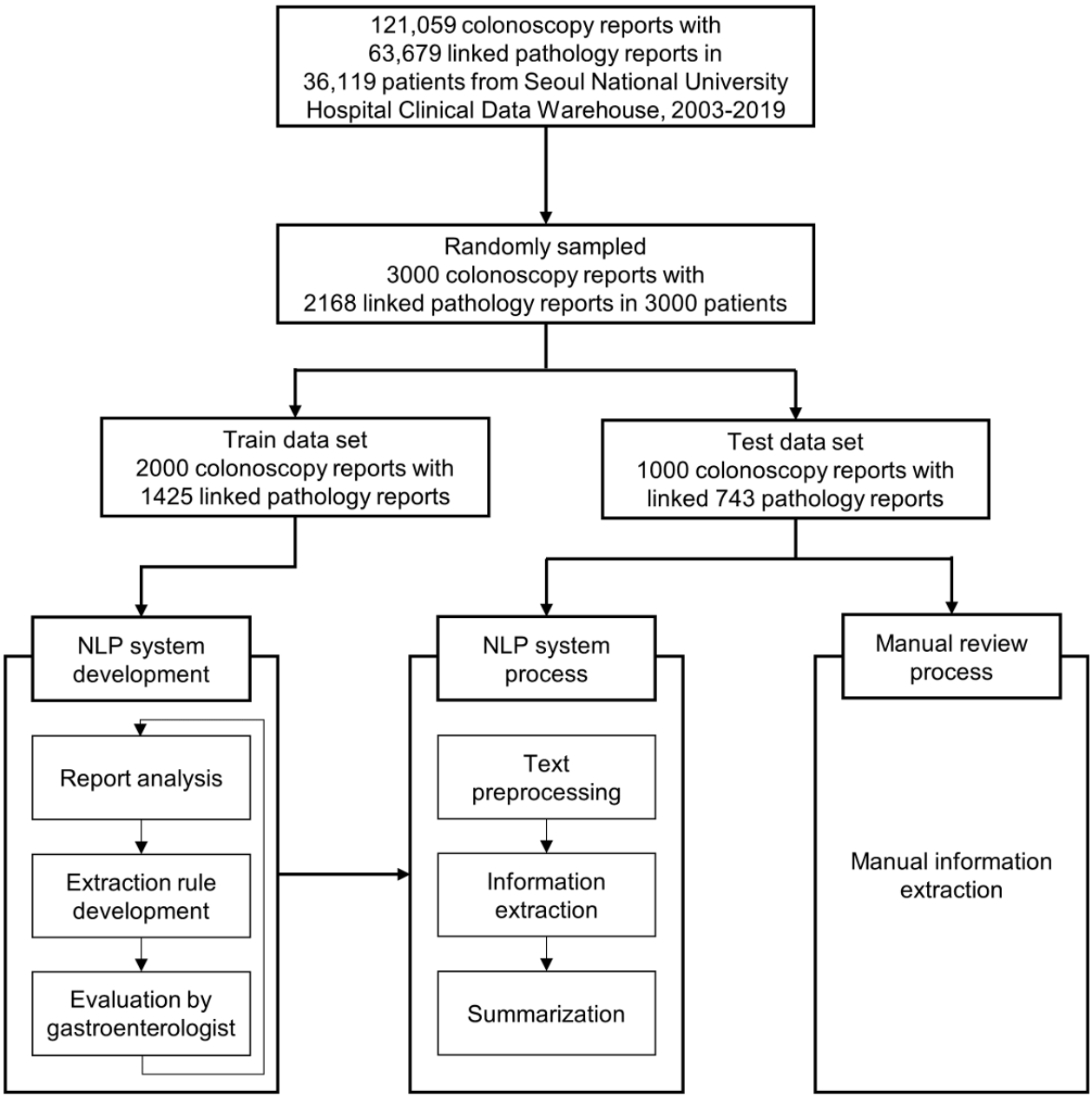
Here, we developed an NLP pipeline for the automated assessment of quality indicators, such as ADR, SDR, and surveillance intervals, from multi-language colonoscopy and pathology report forms. The pipeline was evaluated in a validation set and compared with expert manual reviews to determine whether the pipeline could reliably assist the inefficient manual process. The NLP system was also applied to a 10-year set of colonoscopy and pathology reports to investigate its ability to process real-world data on colonoscopy quality indicators from individual endoscopists.

Methods

Study Design and Population

Colonoscopy for colon cancer screening was performed at Seoul National University Hospital Gangnam Center, where comprehensive medical checkups of approximately 30,000 patients are conducted annually. A total of 121,059 screening and surveillance colonoscopies with 63,697 associated pathology reports from 36,119 patients examined between 2003 and 2019 were derived from SUPREME (Seoul National University Hospital Patients Research Environment), the clinical data warehouse of Seoul National University Hospital. A representative sample of 3000 colonoscopy reports, paired with 2168 pathology reports, from 3000 patients examined after 2003 was randomly selected and used as the development data set for the NLP pipeline (Figure 1). The reports were divided into a training data set of 2000 colonoscopy reports for NLP rule formulation and a testing data set of 1000 colonoscopy reports for validation. Five human annotators (4 board-certified gastroenterologists and 1 researcher) manually reviewed all procedure data and made reference to a consensus of the 5 human annotators for the data set.

Figure 1. Data set description and process for the NLP pipeline development and information extraction. NLP: natural language processing.

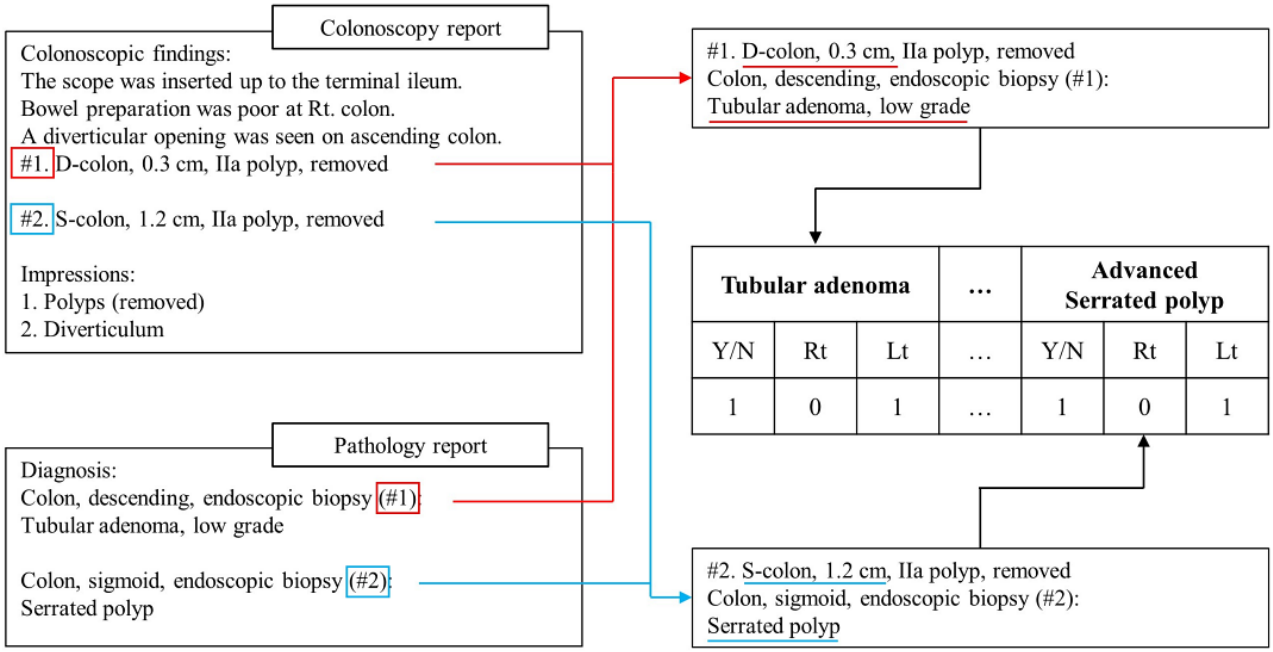


NLP Pipeline Development

We used regular expressions in Python (3.7.10, Python Software Foundation) and smartTA (1.0b, MISO Info Tech) to develop the NLP pipeline. Regular expressions are a sequence of characters specialized for complex text processing using metacharacters [19]. smartTA is NLP software that helps analyze linguistic patterns and construct lexicons. The NLP pipeline was developed with the following steps: First, we developed multi-language report forms (in Korean only, in English only, and a mixed report form) for the NLP pipeline processing by creating a Korean-English lexicon for medical terms, synonyms, and endoscopic abbreviations using a training data set and a colonoscopy textbook [20]. Second, we determined removable terms and phrases in the reports through an interactive discussion with gastroenterologists. Third, we defined the extraction rules using smartTA. Fourth, we updated the rules after the extracted results were evaluated by gastroenterologists. These development steps were repeated until it was no longer possible to obtain performance increases by updating the extraction rules. The final version was validated using the 1000-report testing data set.

The NLP pipeline developed for this study consisted of text preprocessing, information extraction, and summarization (Figure 1, Figure 2). In text preprocessing, the colonoscopy and associated pathology reports were combined as follows: each sentence including a biopsy-related phrase (ie, an abbreviation, number, or character) in the findings section of the colonoscopy report was linked with polyp histopathology results in the diagnosis section of the pathology report according to the sequence of specimens in the pathology report. In information extraction, the pipeline consulted the lexicon to extract the target information, including the presence, type, location, and size of polyps, from the combined colonoscopy-pathology text.

Figure 2. Extraction and summarization process of the NLP pipeline. NLP: natural language process; Y/N: yes/no (indicating presence or absence); Rt: right colon; Lt: left colon.



Finally, the extracted information on the biopsied polyps was summarized in the final summary format and used to calculate the detection rate and surveillance interval.

Target Variables for Polyp Detection and Surveillance Interval Measurement

The NLP tool extracted specific information on colon polyps, such as pathological type, anatomical location, and size. The type of colon polyp was extracted from the pathology reports and categorized as adenoma, serrated polyp, or carcinoma. Additionally, the NLP tool extracted the subcategory for adenomas (ie, tubular, tubulovillous, villous, or adenoma with high-grade dysplasia) and serrated polyps (ie, hyperplastic polyp, SSL, or traditional serrated adenoma). Information on the anatomical location of polyps was extracted from the findings section of the colonoscopy reports and defined as follows: left-colon polyps were defined as those located between the rectum and the splenic flexure (ie, the rectum, rectosigmoid, sigmoid, descending colon, and splenic flexure); right-colon polyps were defined as those located between the transverse colon and the cecum (ie, the transverse colon, hepatic flexure, ascending colon, cecum, and ileocecal valve). When location measurements were provided as the distance from the anal verge in cm, a distance of ≥60 cm was considered to be in the right colon.

The detection rate was calculated as the proportion of colonoscopies that detected at least 1 adenoma or SSL; the overall detection rate and the per-physician detection rate were calculated. The detection rate for advanced adenoma was defined as the proportion of screening colonoscopies that detected a polyp with size ≥1 cm or an adenomatous pathology with high-grade dysplasia or villous features. The detection rate for advanced SSL was defined as the proportion of screening colonoscopies that detected a polyp with a size ≥1 cm or a pathology with low- or high-grade dysplasia. Surveillance intervals were chosen based on the 2020 US Multi-Society Task Force guidelines, which recommend that a patient with neoplastic polyps undergo surveillance colonoscopies at 1 of 6 defined intervals [21].

Statistical Analysis and Performance Evaluation

Continuous variables were calculated as the mean (SD). Discrete data were tabulated as numbers and percentages. The chi-square test was used to compare proportions, and a 2-tailed *t* test was used to compare quantitative variables. Information extraction performance was evaluated by recall, precision, accuracy, and the F1 score. The F1 score is the harmonic mean of precision and recall. Python (3.7.10) and the SciPy package (1.6.2) were used for statistical calculations [22].

Analysis of a 10-Year Set of Colonoscopy Reports for ADR, SDR, and Surveillance Interval

The NLP pipeline analyzed 54,562 screening and surveillance colonoscopy reports and 34,943 associated pathology reports from 12,264 patients aged ≥50 years at Seoul National University Hospital Gangnam Center; all patients were examined between January 2010 and December 2019. The ADR, SDR, and surveillance intervals were investigated, both overall and individually for endoscopists who performed >500 procedures. The relationship between the polyp detection rate and surveillance interval was also determined.

Ethics Approval

This study was approved by the Institutional Review Board of Seoul National University Hospital (1909-093-670).

Results

NLP Information Extraction Performance

Table 1 shows the demographics of the 2000-report training data set and the 1000-report testing data set for the NLP pipeline.

The NLP tool extracted variables to calculate the quality indicators. Table 2 shows the extracted key information on pathological type, including advanced features, location, and the number of polyps, which was assessed for recall, precision, accuracy, and the F1 score in the testing data set. The performance of the NLP pipeline ranged from 0.97 to 1.00 in all performance metrics for the presence of adenomas and SSLs with advanced features. For the location of colon polyps, the NLP pipeline demonstrated excellent performance for adenomas, ranging from 0.97 to 1.00; however, the NLP pipeline demonstrated a relatively lower performance for detecting SSL location. The NLP pipeline also demonstrated high performance (>0.98) for counting the number of adenomas and SSLs.

Table 1. Characteristics of training and testing data sets for the development of the natural language processing pipeline.

Characteristics	Training (N=2000)	Testing (N=1000)	*P* value
Age, mean (SD)	58.6 (6.4)	60.4 (6.5)	<.001
Sex			.86
Male, n (%)	1188 (59.4)	590 (59.0)	
Female, n (%)	812 (40.6)	410 (41.0)	
Adenoma			
Overall, n (%)	925 (46.2)	475 (47.5)	.72
Right colon only, n (%)	501 (25.0)	265 (26.5)	.54
Left colon only, n (%)	212 (10.6)	113 (11.3)	.65
Both, n (%)	212 (10.6)	97 (9.7)	.53
Advanced adenoma[a]			
Overall, n (%)	77 (3.8)	34 (3.4)	.62
Right colon only, n (%)	51 (2.6)	14 (1.4)	.06
Left colon only, n (%)	24 (1.2)	18 (1.8)	.26
Both, n (%)	3 (0.2)	2 (0.2)	.87
Sessile serrated lesion			
Overall, n (%)	121 (6)	66 (6.6)	.64
Right colon only, n (%)	79 (4)	45 (4.5)	.56
Left colon only, n (%)	34 (1.7)	15 (1.5)	.80
Both, n (%)	8 (0.4)	6 (0.6)	.64
Advanced sessile serrated lesion[b]			
Overall, n (%)	19 (1)	12 (1.2)	.66
Right colon only, n (%)	14 (0.7)	10 (1)	.52
Left colon only, n (%)	4 (0.2)	1 (0.1)	.88
Both, n (%)	1 (0.1)	1 (0.1)	.80
Cancer			
Overall, n (%)	3 (0.2)	0 (0)	.54
Right colon only, n (%)	0 (0)	0 (0)	
Left colon only, n (%)	3 (0.2)	0 (0)	.54
Both, n (%)	0 (0)	0 (0)	

[a]Advanced adenomas were defined as adenomas ≥1 cm in size or with pathological features such as high-grade dysplasia or villous features.

[b]Advanced sessile serrated lesions were defined as lesions ≥1 cm in size or with pathological features such as low or high-grade dysplasia.

Table 2. Performance of the natural language processing pipeline in the testing data set (N=1000).

Indicators	Recall	Precision	Accuracy	F1 score
Presence of a conventional adenoma	0.99	1.00	0.99	0.99
Location of conventional adenoma				
None	1.00	0.98	0.99	0.99
Right colon only	0.98	1.00	0.99	0.99
Left colon only	0.98	0.99	0.99	0.99
Both	0.99	0.97	0.99	0.98
Presence of an advanced adenoma[a]	1.00	0.97	0.99	0.99
Location of advanced adenoma				
None	0.99	1.00	0.99	0.99
Right colon only	1.00	0.93	0.99	0.97
Left colon only	1.00	1.00	1.00	1.00
Both	1.00	1.00	1.00	1.00
Presence of an SSL[b]	0.98	1.00	0.99	0.99
Location of SSL				
None	1.00	0.99	0.99	0.99
Right colon only	0.96	1.00	0.99	0.98
Left colon only	1.00	1.00	1.00	1.00
Both	1.00	0.86	0.99	0.92
Presence of an advanced SSL[c]	1.00	1.00	1.00	1.00
Location of advanced SSL				
None	1.00	1.00	1.00	1.00
Right colon only	0.90	1.00	0.99	0.95
Left colon only	1.00	1.00	1.00	1.00
Both	1.00	0.50	0.99	0.67
Total number of adenomas				
0	1.00	0.99	1.00	0.99
1-2	0.99	0.99	0.99	0.99
3-4	0.98	1.00	0.98	0.99
5-10	1.00	1.00	1.00	1.00
>10	N/A[d]	N/A	N/A	N/A
Total number of SSLs				
0	1.00	0.99	1.00	0.99
1-2	0.98	1.00	0.98	0.99
3-4	1.00	1.00	1.00	1.00
5-10	N/A	N/A	N/A	N/A

[a]Advanced adenomas were defined as adenomas ≥1 cm in size or with pathological features such as high-grade dysplasia or villous features.

[b]SSL: sessile serrated lesion.

[c]Advanced sessile serrated lesions were defined as lesions ≥1 cm in size or with pathological features such as low or high-grade dysplasia.

[d]N/A: not applicable.

NLP Performance in Calculating Colonoscopy Quality Indicators

The NLP pipeline assessed the mean ADR and SDR in the test data set as 47.2% (472/1000) and 6.5% (65/1000), respectively. The gold standard evaluation assessed these values as 47.5% (475/1000) and 6.6% (66/1000), respectively (Table 3). The differences in assessed ADR and SDR between the manual review, the NLP pipeline, and the gold standard values were not significant. For assessing the number of patients assigned to each of the 6 surveillance interval groups described in the 2020 US Multi-Society Task Force guidelines, the NLP pipeline and manual review demonstrated similar performance; however, the NLP pipeline demonstrated a relatively higher accuracy in assessing the number of patients assigned to the 3-year group than the manual review (63/63, 100% vs 59/63, 93.6%, respectively); this was also true for the 3-5-year group (68/69, 98.6% vs 65/69, 94.2%, respectively). It is a complicated task to assess risk stratification in these groups.

Table 3. Comparison of polyp detection rate and surveillance interval group assignment as assessed by manual review and the natural language processing pipeline in the test data set (N=1000).

Extracted indicators	Human annotator					Method			*P* value[a]
	A	B	C	D	E	Manual review[b]	NLP system	Gold standard[c]	
Detection rate, n (%)									
ADR[d]	467 (46.7)	474 (47.4)	474 (47.4)	475 (47.5)	468 (46.8)	472 (47.2)	468 (46.8)	475 (47.5)	.92
SDR[e]	65 (6.5)	64 (6.4)	66 (6.6)	64 (6.4)	64 (6.4)	65 (6.5)	64 (6.4)	66 (6.6)	.99
Surveillance interval group, n (%)									
1 year	N/A[f]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3 years	59 (93.7)	58 (92.1)	60 (95.2)	62 (98.4)	58 (92.1)	59 (93.6)	63 (100)	63 (100)	.92
3-5 years	62 (89.9)	67 (97.1)	64 (92.8)	63 (91.3)	68 (98.6)	65 (93.9)	68 (94.2)	69 (100)	.92
5-10 years	40 (100)	40 (100)	40 (100)	40 (100)	40 (100)	40 (100)	39 (97.5)	40 (100)	.99
7-10 years	339 (97.7)	347 (100)	345 (99.4)	345 (99.4)	346 (99.7)	344 (99.1)	343 (98.9)	347 (100)	.99
10 years	479 (99.6)	480 (99.8)	481 (100)	480 (99.8)	480 (99.8)	480 (99.8)	480 (99.8)	481 (100)	.99

[a]*P* values were calculated using the 2X3 chi-square test.

[b]Mean of the judgments made by the 5 human annotators.

[c]Consensus judgment of the 5 human annotators; applied in inconsistent cases.

[d]ADR: adenoma detection rate.

[e]SDR: sessile serrated lesion detection rate.

[f]N/A: not applicable (no patients were assigned a 1-year surveillance interval).

Analysis of ADR, SDR, and Surveillance Intervals in a 10-Year Colonoscopy Report Data Set

The NLP pipeline was applied to a set of 54,562 colonoscopy reports (and their associated pathology reports) created by 25 endoscopists who examined patients aged ≥50 years over a 10-year period; the NLP analyzed ADR, SDR, and surveillance intervals in the reports (Table 4). The overall ADR, advanced ADR, SDR, and advanced SDR were 42% (22,909/54,562), 3.4% (1838/54,562), 3.3% (1806/54,562), and 0.5% (248/54,562), respectively. The difference in detection rate between the endoscopists with the highest and lowest performance was 39.9% (1055/1876, 56.2% vs 264/1615, 16.3%, respectively) for ADR, 5.3% (83/1165, 7.1% vs 30/1615, 1.8%, respectively) for advanced ADR, 6.2% (124/1876, 6.6% vs 6/1615, 0.4%, respectively) for SDR, and 1.6% (11/679, 1.6% vs 0/1615, 0%, respectively) for advanced SDR. Overall, the mean surveillance interval was 8.7 years, and the difference in the surveillance interval assigned by endoscopists with the highest and lowest performance was 1.3 years (9.5 years vs 8.2 years). Table 5 shows the proportion of patients assigned to each of the 6 surveillance interval groups by groups of endoscopists divided according to the endoscopists' ADR and SDR. The group of endoscopists with the lowest ADR (<30%) assigned a higher proportion of patients to the longest surveillance interval than did the endoscopists with the highest ADR (>45%). This pattern was similar for the endoscopists with the highest and lowest SDR.

Table 4. Clinical application of the natural language processing pipeline to nonannotated colonoscopy data created by 25 endoscopists between 2010 and 2019.

Endoscopist	Procedures	Adenoma detection rate, n (%)	Advanced adenoma detection rate, n (%)	Sessile serrated lesion detection rate, n (%)	Advanced sessile serrated lesion detection rate, n (%)	Mean surveillance interval, years
A	3060	1112 (36.3)	94 (3.1)	58 (1.9)	8 (0.3)	8.9
B	981	343 (35)	36 (3.7)	8 (0.8)	0 (0)	9.0
C	3553	1447 (40.7)	129 (3.6)	91 (2.6)	21 (0.6)	8.8
D	2765	1109 (40.1)	92 (3.3)	83 (3)	17 (0.6)	8.8
E	1174	469 (39.9)	46 (3.9)	18 (1.5)	3 (0.3)	8.9
F	1258	338 (26.9)	39 (3.1)	21 (1.7)	1 (0.1)	9.2
G	679	301 (44.3)	12 (1.8)	40 (5.9)	11 (1.6)	8.6
H	1165	505 (43.3)	83 (7.1)	21 (1.8)	4 (0.3)	8.4
I	1615	264 (16.3)	30 (1.9)	6 (0.4)	0 (0)	9.5
J	2091	917 (43.9)	43 (2.1)	92 (4.4)	12 (0.6)	8.7
K	1876	1055 (56.2)	58 (3.1)	124 (6.6)	16 (0.9)	8.2
L	3284	1739 (53)	73 (2.2)	144 (4.4)	14 (0.4)	8.4
M	3437	1510 (43.9)	116 (3.4)	132 (3.8)	3 (0.1)	8.6
N	3799	1708 (45)	119 (3.1)	130 (3.4)	13 (0.3)	8.6
O	647	292 (45.1)	14 (2.2)	14 (2.2)	1 (0.2)	8.8
P	1707	844 (49.4)	74 (4.3)	87 (5.1)	16 (0.9)	8.4
Q	2964	1435 (48.4)	106 (3.6)	137 (4.6)	16 (0.5)	8.5
R	3209	1235 (38.5)	108 (3.4)	99 (3.1)	12 (0.4)	8.8
S	2168	816 (37.6)	52 (2.4)	61 (2.8)	8 (0.4)	8.9
T	3834	1633 (42.6)	119 (3.1)	152 (4)	23 (0.6)	8.7
U	3935	1324 (33.6)	127 (3.2)	68 (1.7)	9 (0.2)	9.1
V	1936	1014 (52.4)	114 (5.9)	104 (5.4)	17 (0.9)	8.2
W	643	268 (41.7)	33 (5.1)	4 (0.6)	0 (0)	8.8
X	1469	680 (46.3)	65 (4.4)	73 (5)	16 (1.1)	8.5
Y	1313	551 (42)	56 (4.3)	39 (3)	7 (0.5)	8.7
Total	54,562	22,909 (42)	1838 (3.4)	1806 (3.3)	248 (0.5)	8.7

Table 5. Proportion of patients assigned different surveillance intervals, sorted by endoscopists (N=25) with high, medium, and low adenoma detection rates and sessile serrated lesion detection rates.

Surveillance interval	Adenoma detection rate, n (%)			Sessile serrated lesion detection rate, n (%)		
	<30% (n=2873)	30%-45% (n=37,806)	>45% (n=13,883)	<2% (n=13,831)	2%-4% (n=24,725)	>4% (n=16,006)
1 year	0 (0)	14 (0.04)	13 (0.09)	3 (0.02)	8 (0.03)	16 (0.1)
3 years	77 (2.68)	1918 (5.07)	894 (6.44)	603 (4.36)	1284 (5.19)	1002 (6.26)
3-5 years	59 (2.05)	2204 (5.83)	1217 (8.77)	545 (3.94)	1557 (6.3)	1378 (8.61)
5-10 years	25 (0.87)	670 (1.77)	389 (2.80)	138 (1.00)	491 (1.99)	455 (2.84)
7-10 years	472 (16.43)	11,213 (29.66)	4953 (35.68)	3527 (25.5)	7508 (30.37)	5603 (35.01)
10 years	2231 (77.75)	21,740 (57.5)	6397 (46.08)	8988 (64.98)	13,851 (56.02)	7529 (47.04)

Discussion

Comparison With Other NLP Systems

There have been various efforts to develop NLP systems for monitoring the quality of colonoscopies in Western countries, and these have shown excellent performance in measuring procedure indications, cecal intubation rate, and the presence and location of polyps. NLP systems have been studied that have various levels of complexity and perform various tasks, ranging from simple extraction tasks, such as assessing the presence and location of polyps, to the automated extraction and calculation of quality metrics [23-31]. However, Western-developed NLP systems in previous studies were based on reports written in English and used NLP lexicons from common language systems, such as the unified medical language system and the Systematized Nomenclature of Medicine-Clinical Terms. These systems cannot be applied to a set of reports written in Korean, both Korean and English, and English only, such as the one examined in this study. Therefore, for the first time in Korea, we developed an NLP pipeline to process colonoscopy reports written in multiple languages. A lexicon including Korean and English medical terms and various endoscopic abbreviations was used to construct the NLP pipeline. Hence, our NLP pipeline processed reports with feasible performance in the validation data set for capturing key quality indicators, including the detection rate for SSLs (previous NLP systems have only captured a few SSLs).

We demonstrated the clinical application of the NLP pipeline with a 10-year set of nonannotated colonoscopy reports. Quality indicators, including ADR, SDR, and surveillance intervals, were extracted from reports written by 25 gastroenterologists, and the proportion of patients assigned different surveillance intervals was analyzed to determine the quality of polyp detection by the endoscopists. We found that ADR and SDR had great variance among the endoscopists, a result that is in line with previous studies [2-4]. There was a 3.4-fold variation in ADR between the endoscopists with the lowest and highest levels (1055/1876, 56.2% vs 264/1615, 16.3%, respectively) and a 16.5-fold variation in SDR (124/1876, 6.6% vs 30/1615, 0.4%, respectively).

Importance of SSL Detection and Performance Feedback

Although awareness of the clinical importance of SSLs for colorectal cancer via the serrated pathway has increased since 2010, our data revealed that detecting SSLs remains a challenge for endoscopists performing screening colonoscopies. SSLs typically show a subtle endoscopic appearance: they can be flat, mucus-coated, and have indistinct borders, which is a totally different appearance from conventional adenomas [32]. Most recently, Lee et al [3] reported the results of a 1-year educational intervention based on a computerized training module that imparted knowledge on the appearance of SSLs using the NICE (Narrow Band Imaging International Colorectal Endoscopic) and WASP (Workgroup on Serrated Polyps and Polyposis) classifications. In this large study, which included 15 experienced endoscopists, the SDR improved significantly, from 4.5% at baseline to 7.1%. Therefore, implementing an NLP system for colonoscopies in clinical practice could provide feedback on the detection performance of individual endoscopists in real time and motivate endoscopists to improve their knowledge and observation techniques for difficult polyps.

Optimization of Surveillance Interval Recommendations

Current surveillance interval recommendations for follow-up colonoscopies do not consider the performance of the physician and only consider the characteristics of the removed polyp. Our study reveals that the recommended surveillance interval can be incorrectly long, depending on the performance level of the endoscopist. High-performance endoscopists (ADR >45%) recommended a 10-year surveillance interval in 46.1% of patients (6397/13,883), while low-performance endoscopists (ADR <30%) recommended a 10-year surveillance interval in 77.8% of patients (2231/2873). This wide difference in the proportion of patients that received a recommendation of a 10-year surveillance interval suggests that low-performance endoscopists missed polyps, negatively affecting their calculation of the future risk of patients and leading them to recommend an inappropriately long surveillance interval. Therefore, endoscopists should periodically check their own ability to detect neoplastic polyps and adjust their recommendations for surveillance interval according to their level of performance to prevent cancer development. Colonoscopy NLP systems could have a role in this self-evaluation process, providing an essential clinical decision support system and enabling the optimal choice of surveillance intervals by considering not only the risk of the patient, but also the performance of the endoscopist.

Limitations

This study has the following limitations: First, it was conducted at a single center, leaving open the possibility that the NLP pipeline may not be able to properly process colonoscopy reports retrieved from other centers. As the NLP pipeline is based on regular expression rules formulated from linguistic patterns in the development data set, terms or patterns in other reports that are not present in the development data set can result in false processing of the reports. Second, the integrity of the NLP pipeline depends on the endoscopist's documentation practice. For example, miswriting orders, numbers, or the count of the biopsied polyps could create mismatches between a colonoscopy report and its associated pathology report, resulting in false processing in the pipeline. However, this is not a problem unique to our study; it applies to all projects that use current NLP pipelines. Therefore, future research may be required to develop more confident NLP systems that warn of the possibility of false processing or to develop more sophisticated systems based on deep learning approaches and cutting-edge NLP models, such as bidirectional encoder representations from transformers (BERT) [33].

Conclusions

In summary, we developed an NLP pipeline to transform multi-language, free-text reports into a structured format to automate the calculation of quality indicators. The NLP pipeline processed the validation data set with high performance that

was similar to a manual review performed by experts. The NLP-derived information from a 10-year real-world data set found that individual endoscopists showed great variance in quality indicators and patient risk stratification. This automated NLP process could be a useful decision support system for endoscopists, as it could allow the optimal recommendation of postcolonoscopy surveillance intervals based on both patient risk and endoscopist performance. This system could positively impact the quality of colonoscopy in many hospitals and health check-up centers that conduct screening programs. Furthermore, information extracted by NLP pipelines from big data derived from colonoscopy reports should be a valuable resource for research into the association of colon polyps with various diseases and into guideline adherence patterns.

Acknowledgments

This study was supported by a National Research Foundation of Korea grant funded by the Korean government (grants 2019R1F1A1061665 and 2020R1F1A1068423).

Authors' Contributions

JHB contributed to conceptualization. SS, SYY, JYS, EHJ, GEC, SJC, HCK, and DUA contributed to data collection and material preparation. JHB contributed to the formal analysis. GS and JHB contributed to devising the methodology. GS and JHB wrote and prepared the original draft. HWH and SYY contributed to writing, reviewing, and editing. HWH and SYY were supervisors.

Conflicts of Interest

None declared.

References

1. Senore C, Basu P, Anttila A, Ponti A, Tomatis M, Vale DB, et al. Performance of colorectal cancer screening in the European Union Member States: data from the second European screening report. Gut 2019 Jul 10;68(7):1232-1244. [doi: 10.1136/gutjnl-2018-317293] [Medline: 30530530]
2. Burr NE, Derbyshire E, Taylor J, Whalley S, Subramanian V, Finan PJ, et al. Variation in post-colonoscopy colorectal cancer across colonoscopy providers in English National Health Service: population based cohort study. BMJ 2019 Nov 13;367:l6090 [FREE Full text] [doi: 10.1136/bmj.l6090] [Medline: 31722875]
3. Lee J, Bae JH, Chung SJ, Kang HY, Kang SJ, Kwak M, et al. Impact of comprehensive optical diagnosis training using Workgroup serrAted polypS and Polyposis classification on detection of adenoma and sessile serrated lesion. Dig Endosc 2022 Jan 12;34(1):180-190. [doi: 10.1111/den.14046] [Medline: 34021513]
4. Hetzel J, Huang C, Coukos J, Omstead K, Cerda SR, Yang S, et al. Variation in the detection of serrated polyps in an average risk colorectal cancer screening cohort. Am J Gastroenterol 2010 Dec;105(12):2656-2664. [doi: 10.1038/ajg.2010.315] [Medline: 20717107]
5. Rex DK, Schoenfeld PS, Cohen J, Pike IM, Adler DG, Fennerty MB, et al. Quality indicators for colonoscopy. Gastrointest Endosc 2015 Jan;81(1):31-53. [doi: 10.1016/j.gie.2014.07.058] [Medline: 25480100]
6. Anderson JC, Butterly LF, Weiss JE, Robinson CM. Providing data for serrated polyp detection rate benchmarks: an analysis of the New Hampshire Colonoscopy Registry. Gastrointest Endosc 2017 Jun;85(6):1188-1194 [FREE Full text] [doi: 10.1016/j.gie.2017.01.020] [Medline: 28153571]
7. Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 2014 Apr 03;370(14):1298-1306 [FREE Full text] [doi: 10.1056/NEJMoa1309086] [Medline: 24693890]
8. Lieberman DA, Rex DK, Winawer SJ, Giardiello FM, Johnson DA, Levin TR. Guidelines for colonoscopy surveillance after screening and polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2012 Sep;143(3):844-857. [doi: 10.1053/j.gastro.2012.06.001] [Medline: 22763141]
9. Lieberman D, Nadel M, Smith RA, Atkin W, Duggirala SB, Fletcher R, et al. Standardized colonoscopy reporting and data system: report of the Quality Assurance Task Group of the National Colorectal Cancer Roundtable. Gastrointest Endosc 2007 May;65(6):757-766. [doi: 10.1016/j.gie.2006.12.055] [Medline: 17466195]
10. Abdul-Baki H, Schoen RE, Dean K, Rose S, Leffler DA, Kuganeswaran E, et al. Public reporting of colonoscopy quality is associated with an increase in endoscopist adenoma detection rate. Gastrointest Endosc 2015 Oct;82(4):676-682 [FREE Full text] [doi: 10.1016/j.gie.2014.12.058] [Medline: 26385276]
11. Sey M, Liu A, Asfaha S, Siebring V, Jairath V, Yan B. Performance report cards increase adenoma detection rate. Endosc Int Open 2017 Jul 06;5(7):E675-E682 [FREE Full text] [doi: 10.1055/s-0043-110568] [Medline: 28691053]
12. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
13. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]

14. Nehme F, Feldman K. Evolving Role and Future Directions of Natural Language Processing in Gastroenterology. Dig Dis Sci 2021 Jan 27;66(1):29-40. [doi: 10.1007/s10620-020-06156-y] [Medline: 32107677]
15. Ridgway JP, Uvin A, Schmitt J, Oliwa T, Almirol E, Devlin S, et al. Natural Language Processing of Clinical Notes to Identify Mental Illness and Substance Use Among People Living with HIV: Retrospective Cohort Study. JMIR Med Inform 2021 Mar 10;9(3):e23456 [FREE Full text] [doi: 10.2196/23456] [Medline: 33688848]
16. Arnaud É, Elbattah M, Gignon M, Dequen G. Deep Learning to Predict Hospitalization at Triage: Integration of Structured Data and Unstructured Text. : IEEE; 2021 Presented at: IEEE International Conference on Big Data; Dec 10-13, 2020; Atlanta, GA URL: https://ieeexplore.ieee.org/abstract/document/9378073 [doi: 10.1109/BigData50022.2020.9378073]
17. Levis M, Leonard Westgate C, Gui J, Watts BV, Shiner B. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychol. Med 2020 Feb 17;51(8):1382-1391. [doi: 10.1017/s0033291720000173]
18. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
19. Van Rossum G. The Python Library Reference, release 3.7.10. In: The Python Library Reference. Amsterdam, The Netherlands: Python Software Foundation; 2020.
20. Song GA, Kim SH, Kim TO, Kang DH. Guide of gastroenterological endoscopy in clinical practice: Korean Society of Gastrointestinal Endoscopy. Seoul, Republic of Korea: 대한의학서적; Jul 03, 2013.
21. Gupta S, Lieberman D, Anderson JC, Burke CA, Dominitz JA, Kaltenbach T, et al. Recommendations for Follow-Up After Colonoscopy and Polypectomy: A Consensus Update by the US Multi-Society Task Force on Colorectal Cancer. Am J Gastroenterol 2020 Mar 7;115(3):415-434 [FREE Full text] [doi: 10.14309/ajg.0000000000000544] [Medline: 32039982]
22. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar 3;17(3):261-272 [FREE Full text] [doi: 10.1038/s41592-019-0686-2] [Medline: 32015543]
23. Deutsch JC. Colonoscopy quality, quality measures, and a natural language processing tool for electronic health records. Gastrointest Endosc 2012 Jun;75(6):1240-1242. [doi: 10.1016/j.gie.2012.02.031] [Medline: 22624812]
24. Gawron AJ, Thompson WK, Keswani RN, Rasmussen LV, Kho AN. Anatomic and advanced adenoma detection rates as quality metrics determined via natural language processing. Am J Gastroenterol 2014 Dec;109(12):1844-1849. [doi: 10.1038/ajg.2014.147] [Medline: 24935271]
25. Harkema H, Chapman WW, Saul M, Dellon ES, Schoen RE, Mehrotra A. Developing a natural language processing application for measuring the quality of colonoscopy procedures. J Am Med Inform Assoc 2011 Dec;18 Suppl 1:i150-i156 [FREE Full text] [doi: 10.1136/amiajnl-2011-000431] [Medline: 21946240]
26. Imler TD, Morea J, Imperiale TF. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. Clin Gastroenterol Hepatol 2014 Jul;12(7):1130-1136. [doi: 10.1016/j.cgh.2013.11.025] [Medline: 24316106]
27. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol 2013 Jun;11(6):689-694 [FREE Full text] [doi: 10.1016/j.cgh.2012.11.035] [Medline: 23313839]
28. Imler TD, Morea J, Kahi C, Sherer EA, Cardwell J, Johnson CS, et al. Multi-center colonoscopy quality measurement utilizing natural language processing. Am J Gastroenterol 2015 Apr;110(4):543-552. [doi: 10.1038/ajg.2015.51] [Medline: 25756240]
29. Lee JK, Jensen CD, Levin TR, Zauber AG, Doubeni CA, Zhao WK, et al. Accurate Identification of Colonoscopy Quality and Polyp Findings Using Natural Language Processing. J Clin Gastroenterol 2019 Jan;53(1):e25-e30 [FREE Full text] [doi: 10.1097/MCG.0000000000000929] [Medline: 28906424]
30. Mehrotra A, Dellon ES, Schoen RE, Saul M, Bishehsari F, Farmer C, et al. Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. Gastrointest Endosc 2012 Jun;75(6):1233-9.e14 [FREE Full text] [doi: 10.1016/j.gie.2012.01.045] [Medline: 22482913]
31. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015 Sep;82(3):512-519 [FREE Full text] [doi: 10.1016/j.gie.2015.01.049] [Medline: 25910665]
32. Musquer N. IJspeert J, Bastiaansen B, Leerdam M, et al (2016) Development and validation of the WASP classification system for optical diagnosis of adenomas, hyperplastic polyps and sessile serrated adenomas/polyps. Gut 65:963–970. Colon Rectum 2018 Aug 21;12(3):200-203. [doi: 10.3166/cer-2018-0029]
33. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019 Presented at: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2-7 June. 2019; Minneapolis, Minnesota p. 4171-4186.

Abbreviations

ADR: adenoma detection rate
NLP: natural language processing
SDR: sessile serrated lesion detection rate
SSL: sessile serrated lesion

Edited by C Lovis; submitted 29.11.21; peer-reviewed by M Elbattah; comments to author 20.12.21; revised version received 13.02.22; accepted 25.02.22; published 15.04.22.

Please cite as:
Bae JH, Han HW, Yang SY, Song G, Sa S, Chung GE, Seo JY, Jin EH, Kim H, An D
Natural Language Processing for Assessing Quality Indicators in Free-Text Colonoscopy and Pathology Reports: Development and Usability Study
JMIR Med Inform 2022;10(4):e35257
URL: https://medinform.jmir.org/2022/4/e35257
doi:10.2196/35257
PMID:35436226

Original Paper

Multi-Label Classification in Patient-Doctor Dialogues With the RoBERTa-WWM-ext + CNN (Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach With Whole Word Masking Extended Combining a Convolutional Neural Network) Model: Named Entity Study

Yuanyuan Sun[1,2], BSc; Dongping Gao[1], PhD; Xifeng Shen[1], BSc; Meiting Li[1], BSc; Jiale Nan[1], BA; Weining Zhang[1], BSc

[1]Institute of Medical Information, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China

[2]Department of Internal Medicine, Chinese Academy of Medical Sciences, Peking Union Medical College Hospital, Beijing, China

Corresponding Author:
Dongping Gao, PhD
Institute of Medical Information
Chinese Academy of Medical Sciences
Peking Union Medical College
No 3 Yabao Road
Chaoyang District
Beijing, 100020
China
Phone: 86 10 5232 8720
Email: gaodp_gaodp@126.com

Abstract

Background: With the prevalence of online consultation, many patient-doctor dialogues have accumulated, which, in an authentic language environment, are of significant value to the research and development of intelligent question answering and automated triage in recent natural language processing studies.

Objective: The purpose of this study was to design a front-end task module for the network inquiry of intelligent medical services. Through the study of automatic labeling of real doctor-patient dialogue text on the internet, a method of identifying the negative and positive entities of dialogues with higher accuracy has been explored.

Methods: The data set used for this study was from the Spring Rain Doctor internet online consultation, which was downloaded from the official data set of Alibaba Tianchi Lab. We proposed a composite abutting joint model, which was able to automatically classify the types of clinical finding entities into the following 4 attributes: positive, negative, other, and empty. We adapted a downstream architecture in Chinese Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach (RoBERTa) with whole word masking (WWM) extended (RoBERTa-WWM-ext) combining a text convolutional neural network (CNN). We used RoBERTa-WWM-ext to express sentence semantics as a text vector and then extracted the local features of the sentence through the CNN, which was our new fusion model. To verify its knowledge learning ability, we chose Enhanced Representation through Knowledge Integration (ERNIE), original Bidirectional Encoder Representations from Transformers (BERT), and Chinese BERT with WWM to perform the same task, and then compared the results. Precision, recall, and macro-F1 were used to evaluate the performance of the methods.

Results: We found that the ERNIE model, which was trained with a large Chinese corpus, had a total score (macro-F1) of 65.78290014, while BERT and BERT-WWM had scores of 53.18247117 and 69.2795315, respectively. Our composite abutting joint model (RoBERTa-WWM-ext + CNN) had a macro-F1 value of 70.55936311, showing that our model outperformed the other models in the task.

Conclusions: The accuracy of the original model can be greatly improved by giving priority to WWM and replacing the word-based mask with unit to classify and label medical entities. Better results can be obtained by effectively optimizing the

downstream tasks of the model and the integration of multiple models later on. The study findings contribute to the translation of online consultation information into machine-readable information.

(JMIR Med Inform 2022;10(4):e35606) doi:10.2196/35606

KEYWORDS

online consultation; named entity; automatic classification; ERNIE; Enhanced Representation through Knowledge Integration; BERT; Bidirectional Encoder Representations from Transformers; machine learning; neural network; model; China; Chinese; classification; patient-physician dialogue; patient doctor dialogue; semantics; natural language processing

Introduction

Background

Internet hospitals in China are in high demand due to limited and unevenly distributed health care resources, lack of family physicians, increasing burden of chronic diseases, and rapid growth of the aging population [1]. Gong et al researched online epidemic-related consultations by multicenter internet hospitals in China during the COVID-19 epidemic, and proved that internet hospitals can offer essential medical support to the public, reduce social panic, and reduce the chance of nosocomial cross-infection, thus playing an important role in preventing and controlling COVID-19 [2]. The COVID-19 outbreak catalyzed the expansion of online health care services. During online consultation, large amounts of text data are accumulated, and contextual data that contain patient-doctor dialogues are of significant value. Network inquiry technology is still in the popularization stage in China, and the text record of inquiry is seldom used in research in the area of natural language processing (NLP), which involves patient privacy and information security [3]. Recently, there has been a lot of work in this area, for instance, a study on the problem of corpus-level entity typing [4]. Chinese scholars have reported on multi-instance learning in the 27th ACM International Conference [5]. Moreover, Wentong et al introduced named entity recognition of electronic medical records based on Bidirectional Encoder Representations from Transformers (BERT) [6] and Piao et al researched a Chinese named entity recognition method based on BERT embedding, which improved entity recognition and attribute labeling [7]. These are significant studies in the NLP domain. Entity studies of clinical text data commonly involve electronic medical records. Dun-Wei et al performed a study based on multi-feature embedding and the attention mechanism [8], and Xue et al researched cross-department chunking [9]. Moreover, Zhang et al studied automatic identification of Chinese clinical entities from free text in electronic health records and contributed to translating human-readable health information into machine-readable information [10]. Furthermore, Jiang et al used machine learning approaches to mine massive service data from the largest China-based online medical consultation platform, which covers 1,582,564 consultation records of patient-physician pairs from 2009 to 2018, and showed that promoting multiple timely responses in patient-provider interactions is essential to encourage payment [11].

However, there is limited clinical dialogue data, and the development of sentence compression for aspect-based sentiment analysis is constantly improving [12]. Chinese researchers have used the BERT model to analyze public emotion during the epidemic of COVID-19 and have substantiated that the fine-tuning of BERT has higher accuracy in the training process [13]. A team from Drexel University used a transformer-based machine learning model to analyze the nuances of vaccine sentiment in Twitter discourse [14]. Patient-doctor dialogues, which are different from daily communication or other universal Q&A, contain important data, such as a patient's symptoms and the diagnosis by a doctor, and these are called "clinical findings" or named entities in patient-doctor dialogues.

Objectives

The purpose of this study was to design a front-end task module for the network inquiry of Intelligent Medical Services. Through the study of automatic labeling of real doctor-patient dialogue text on the internet, a method of identifying the negative and positive entities of the dialogue with higher accuracy was explored. This work significantly eliminates the human work involved in feature engineering.

Methods

Data Sets

In this paper, our task was named entity automatic classification in patient-doctor dialogues, which was divided into the following 4 attributes: positive, negative, other, and empty. The details are presented below.

The tag "positive (POS)" is used when it can be determined that a patient has dependent symptoms, diseases, and corresponding entities that are likely to cause a certain disease. The tag "negative (NEG)" is used when the disease and symptoms are not related. The tag "other (OTHER)" is used when the user does not know or the answer is unclear/ambiguous, which is difficult to infer. The tag "empty (EMPTY)" is used when there is no practical meaning to determine the patient's condition, such as interpretation of some medical knowledge by the doctor, independent of the patient's current condition, inspection items, drug names, etc.

The data set is from the *Spring Rain Doctor* internet online consultation, which has been downloaded from the official data set of Alibaba Tianchi Lab [15]. The training set consists of 6000 dialogues, and each set of dialogues contains more than a dozen statements and a total of 186,305 sentences. The test set consists of 2000 dialogues and a total of 61,207 sentences.

On analysis, we found that online consultation data had the below features.

1. The patient description information was scattered, had slang, and had some spelling mistakes:

患者：经常放屁，很丑(臭) (sentence_id:20); Patient: Fart often. It stnks (stinks)

医生：杀菌治疗的话应该重新换药 (sentence_id:21); Doctor: For bactericidal treatment, you should be replaced with drugs

患者：现在安(按)肚脐左边，感觉按着涨涨的感觉 (sentence_id:22); Patient: Now prress (press) the left side of the navel, I feel it like a balloon.

医生：我觉得这种疼痛应该有中药的影响。(sentence_id:23); Doctor: I think this pain should be affected by traditional Chinese medicine.

2. Interval answers were common:

医生：咳嗽咳痰？(sentence_id:4); Doctor: Any Cough or expectoration?

医生：头痛头晕脑胀？(sentence_id:5); Doctor: Headache, dizziness, or brain swelling?

医生：从资料分析看，有可能是过敏性鼻炎。(sentence_id:6); Doctor: According to the previous examination, it may be allergic rhinitis.

患者：应该是里面，表面上没有鼓包或红肿之类的，没有感冒或咳嗽过最近，头晕脑胀有时会 (sentence_id:7); Patient: It should be inside. There is no bulge or swelling on the surface. There is no cold or cough recently. Dizziness and brain swelling sometimes occur.

3. The main symptoms were mixed with other symptoms:

医生：你好，是10岁的孩子头痛吗？(sentence_id:2); Doctor: Hello, is it a 10-year-old child with a *headache*?

患者：是的 (sentence_id:3); Patient: Yes

患者：不知道头疼恶心吐，是不是感冒 (sentence_id:19); Patient: I'm not sure whether headache, *nausea*, or *vomiting* is *colds*

医生：但是感冒一般不会呕吐 (sentence_id:28); Doctor: But a cold usually doesn't cause vomiting

患者：恶心之前没劲，反酸水 (sentence_id:30); Patient: *No strength* before nausea, *sour stomach*

医生：需要详细的问诊和查体，建议到医院神经内科或儿童神经内科面诊 (sentence_id:36); Doctor: Need detailed consultation and physical examination, I suggest going to the hospital *neurology department* or children's neurology department for a face-to-face diagnosis

The above aspects introduce many difficulties in entity recognition and attribute annotation.

The format of raw data was multilayer nested JSON. According to the aspects of the models, we split the innermost text into pairs of splicing contextual sentences. "Jsonlite" is a unique package of R language [16], and the built-in "stream_in" statement does well with tiling JSON into an Excel table, making it intuitive and convenient for us to compare the differences in output data. We then extracted the corresponding subform data according to the analysis requirements. All models shared the same data set. Before input into our model, in addition to the sentence content, we appended the speech role information (ie, sender).

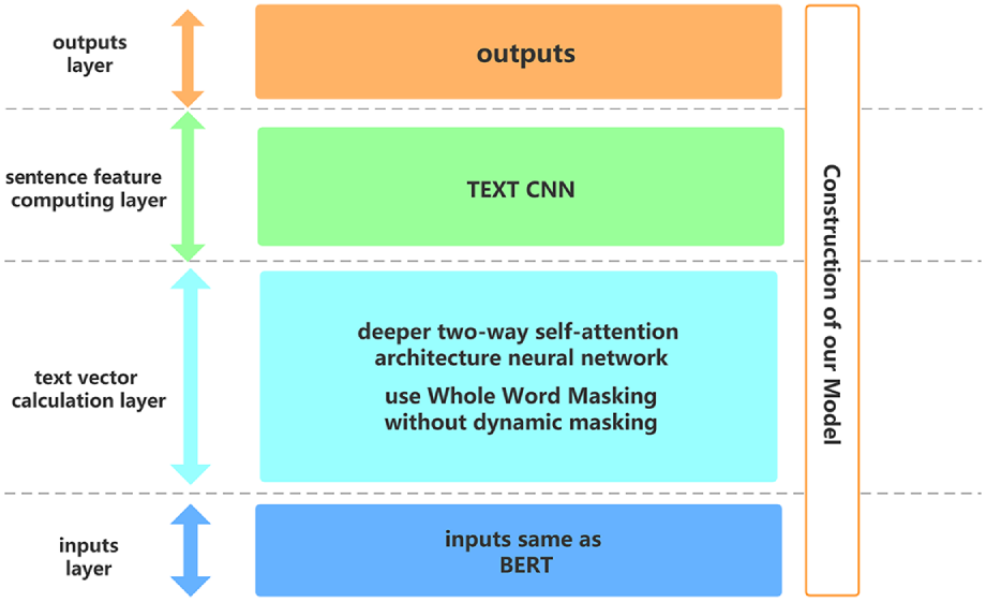
Composite Abutting Joint Model for Clinical Named Entity Classification

We proposed a composite abutting joint model and adapted a downstream architecture in Chinese Encoder Representations from Transformers Pretraining Approach (RoBERTa) with whole word masking (WWM) extended (RoBERTa-WWM-ext), which combines a text convolutional neural network (CNN) [17]. We used RoBERTa-WWM-ext to express sentence semantics as a text vector [18] and then extracted the local features of the sentence through the CNN, which was our new fusion model.

Construction of the Composite Abutting Joint Model

Chinese RoBERTa-WWM-ext is an open-source model from the Harbin Institute of Technology, which uses WWM combined with the RoBERTa model [19,20]. We adapted a downstream architecture in Chinese RoBERTa-WWM, which combines a text CNN [21]. Our training objective was to use RoBERTa-WWM-ext to express sentence semantics as a text vector and then extract the local features of the sentence through the CNN. The construction of our model is shown in Figure 1.

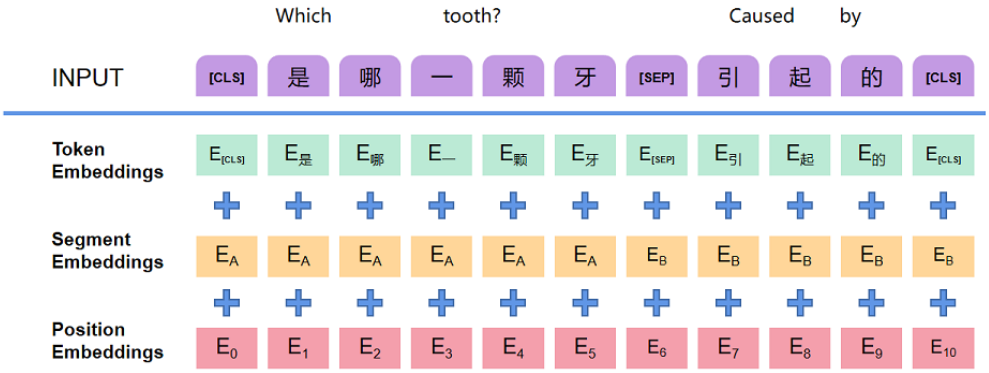
Figure 1. Construction of our model. BERT: Bidirectional Encoder Representations from Transformers; CNN: convolutional neural network.



The Input Layer of the Composite Abutting Joint Model

The input layer is the same as BERT [22]. It uses a masked language model (MLM) to generate deep 2-way linguistic representations that combine adjacent and contextual information. Its structure involves stacking traditional transformers, and taking BERT as an example, each of its 12 transformer layers combine left and right contexts to form a deeper 2-way self-attention architecture neural network. Text-input BERT is characterized by 3 levels (Figure 2), namely, token embeddings, segment embeddings, and position embeddings.

Figure 2. Bidirectional Encoder Representations from Transformers input characterization.



Text Vector Calculation Layer of the Composite Abutting Joint Model

To maintain continuity between sentences, the beginning and end of the original text are marked with a special symbol [CLS], and the 2 sentences are split with [SEP]. The coded information in the discrete state is transformed into N-dimensional space vectors and transmitted to the encoder unit of the transformer through a continuous and distributed representation. Similarity and distance are computed at the self-attention level to capture word dependencies within sentences. For the calculation of the self-attention function, Vaswani et al introduced "Scaled Dot-Product Attention" [23]. The input includes queries and keys for dimension d_k and the value for dimension d_v. The dot products of a query are computed with all keys, and each is divided by each key. Then, the softmax function is applied to the values. In fact, during the model computation, it has a set of queries packed together into a matrix Q. The keys and values are packed together into matrices K and V. The output matrix is as follows [23]:

The model could project the queries, keys, and values linearly h times with different learned linear projections to d_k, d_k, and d_v dimensions, respectively. On each projected version of the queries, keys, and values, it executes the attention function in parallel to generate d_v-dimensional output values. These values are connected and projected again to obtain the final result. This is multihead attention [23].

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \quad \textbf{(2)}$$

where $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ and where the projections are parameter matrices , ,  , and .

The inputs and outputs of the self-attention layer are added and normalized, which makes the output mean of the self-attention layer 0 and the standard deviation 1, and then, it is transferred to the feed-forward layer of the feed-forward neural network. Mean and normalization are processed again. The transformer encoder structure of the model has been described by Vaswani et al [23] (Figure 3).

In transformers, location coding is computed using a trigonometric function as follows [23]:

The positional encoding vector results are added to the embedding vector sequence corresponding to each input word instead of connecting vector. Similar to BERT in our model, 15% of the word-piece tokens are masked at random during training. These masked tokens are divided into 3 parts, with 80% of them using [MASK], 10% of them being replaced with a random word, and 10% of them using the original word. Related research by Dandan et al showed that the downstream task of the pretraining model can improve the performance of the model through FINETUNE [24].

During the pretraining phase, the BERT model takes on 2 tasks, MLM and next sentence prediction (NSP). Piao et al have explained the process of predictive masking in MLM tasks, which obtains the semantic representation of a word in a specific context through self-supervised learning [7]. Not the same as BERT, RoBERTa-WWM-ext cancels the NSP and uses max_len = 512 during the pretraining, and the number of training steps is appropriately extended [18].

Another feature of RoBERTa-WWM-ext is that it uses WWM. An example to illustrate the characteristics of WWM is provided in Figure 4 [19].

BERT can only divide Chinese into characters, not words (units). WWM makes the Chinese mask more like English. A complete word will be shielded; otherwise, it will not be shielded, which can maintain the integrity of the Chinese word as a unit, to improve the accuracy of model learning.

Figure 3. Transformer encoder structure.

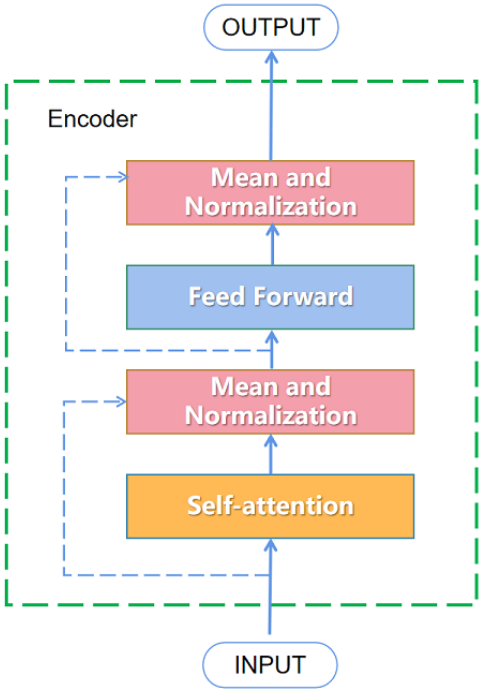
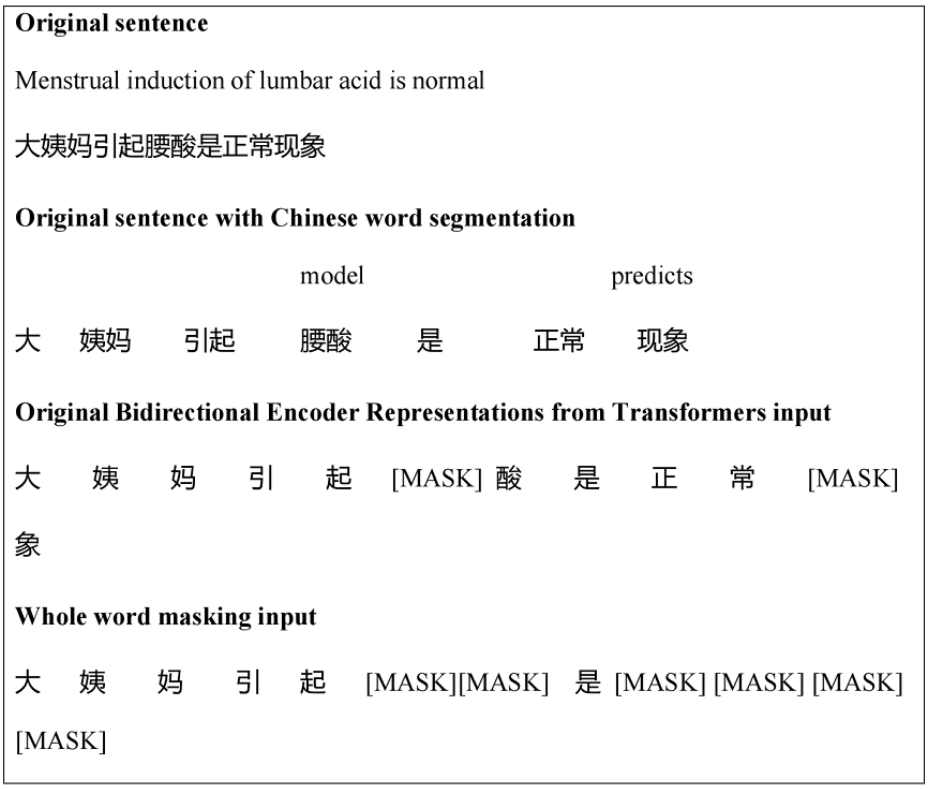


Figure 4. An example of whole word masking in our model.



Sentence Feature Computing Layer of the Composite Abutting Joint Model

The output word vector of RoBERTa-WWM-ext was further extracted by a CNN, which is expected to enhance the robustness of the model. The computing formula is as follows [17,25,26]:

where W_A and W_B are 2 matrices that are randomly initialized by adding an attention layer to deal with the location characteristics, and b is the RoBERTa-WWM-ext hidden layer dimension, with b_1 being the offset. Moreover, E_{Ro} represents the output of the coding layer of RoBERTa-WWM-ext, and $feature_{text}$ represents the weighted feature obtained by the product of the score weight and the output of the encoder, which is also the output text vector feature of RoBERTa-WWM-ext. After CNN calculation, the predicted emotion label is finally obtained [27].

Results

Evaluation Criteria

We adopted Alibaba cloud's official evaluation standard, and Macro-F1 was used as the evaluation index. Suppose we have n categories, C1, ..., CI, ..., CN, the calculation is as follows:

where accuracy (Pi) is the number of samples correctly predicted as category CI/number of samples predicted as category CI, and recall rate (Ri) is the number of samples correctly predicted as category CI/number of samples of the real CI category.

Graphics Processing Unit Server Requirements

The server requirements are as follows: CPU, 8 cores at 2.5 GHz; memory, 32 GB; hard disk, 500 GB; GPU/Field Programmable Gate Array, 1×NVIDIA V100.

Results of Our Composite Abutting Joint Model

Our data involved a 3-layer nested JSON file. The first layer was regarded as the index of each dialogue, the second layer was the specific dialogue content between patients and doctors in each dialogue, and the third layer was the entity part corresponding to a single sentence. Not every sentence had an entity part, and not every entity needed to be marked with an entity attribute. We expanded the training set data and all the models' training results. The distribution of entity attribute labels is shown in Table 1.

From Table 1, we know that the BERT results of the test data have more positive labels, with a value nearly 10 percentage points higher than that for the train data, and the negative labels were nearly 4 percentage points less than that for the train data. After optimizing WWM, the attribute proportion was close to the train data, but there was still a certain gap. We used the fine-tune approach with CNN for RoBERTa-WWM-ext, but it did not change the label proportion. In the Enhanced Representation through Knowledge Integration (ERNIE) model train results, the attribute proportion was closer to that for the

train data when compared with BERT. Next, we compared the 4 models, and the results are shown in Table 2.

Table 1. Attribute statistics in the training data and the model training results of the test data.

Data set	Training data (N=118,976)	Test data (N=39,204)			
		ERNIE[a]	BERT[b]	BERT-WWM[c]	RoBERTa-WWM-ext + CNN[d]
POS[e], n (%)	74,774 (62.85%)	25,163 (64.18%)	27,866 (71.08%)	26,116 (66.62%)	26,116 (66.62%)
NEG[f], n (%)	14,086 (11.84%)	4271 (10.89%)	3125 (7.97%)	3871 (9.87%)	3871 (9.87%)
OTHER[g], n (%)	6167 (5.18%)	1006 (2.57%)	684 (1.74%)	2587 (6.60%)	2587 (6.60%)
EMPTY[h], n (%)	23,949 (20.13%)	8764 (22.35%)	7529 (19.20%)	6630 (16.91%)	6630 (16.91%)

[a]ERNIE: Enhanced Representation through Knowledge Integration.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]BERT-WWM: Bidirectional Encoder Representations from Transformers with whole word masking.

[d]RoBERTa-WWM-ext + CNN: Robustly Optimized BERT Pretraining Approach with whole word masking extended plus a convolutional neural network.

[e]The tag “positive (POS)” is used when it can be determined that a patient has dependent symptoms, diseases, and corresponding entities that are likely to cause a certain disease.

[f]NEG: The tag “negative (NEG)” is used when the disease and symptoms are not related.

[g]OTHER: The tag “other (OTHER)” is used when the user does not know or the answer is unclear/ambiguous, which is difficult to infer.

[h]EMPTY: The tag “empty (EMPTY)” is used when there is no practical meaning to determine the patient’s condition, such as interpretation of some medical knowledge by the doctor, independent of the patient’s current condition, inspection items, drug names, etc.

Table 2. The scores of the 4 models.

Data set	ERNIE[a]	BERT[b] BERT	BERT-WWM[c]	RoBERTa-WWM-ext + CNN[d]
POS[e]-Rr[f]	87.32461545	87.10998052	89.81676537	89.23248142
POS-Pr	87.35933834	78.69582391	86.57854406	88.20871479
POS-F1	87.34197344	82.68940537	88.16793149	88.71764473
NEG[g]-Rr[h]	67.70158588	41.50100514	66.96448515	70.13625195
NEG-Pr	71.03351301	59.45600000	77.50775595	77.30182176
NEG-F1	69.32753888	48.88187319	71.85140803	73.54491158
OTHER[i]-Rr	27.30551262	12.98299845	58.06285420	57.13549717
OTHER-Pr	52.68389662	36.84210526	43.58081980	45.06298253
OTHER-F1	35.96878181	19.20000000	49.79014800	50.38618810
EMPTY[j]-Rr	75.84846093	61.62851881	62.98342541	67.77163904
EMPTY-Pr	65.84446728	62.29224837	72.27169811	71.50589868
EMPTY-F1	70.49330644	61.95860610	67.30863850	69.58870804
Macro-Rr	64.54504372	50.80562573	69.45688253	71.06896740
Macro-Pr	69.23030381	59.32154439	69.98470448	70.51985444
Total score (Macro-F1)	65.78290014	53.18247117	69.27953150	70.55936311

[a]ERNIE: Enhanced Representation through Knowledge Integration.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]BERT-WWM: Bidirectional Encoder Representations from Transformers with whole word masking.

[d]RoBERTa-WWM-ext + CNN: Robustly Optimized BERT Pretraining Approach with whole word masking extended plus a convolutional neural network.

[e]The tag "positive (POS)" is used when it can be determined that a patient has dependent symptoms, diseases, and corresponding entities that are likely to cause a certain disease.

[f]Pr: precision rate.

[g]NEG: The tag "negative (NEG)" is used when the disease and symptoms are not related.

[h]Rr: recall rate.

[i]OTHER: The tag "other (OTHER)" is used when the user does not know or the answer is unclear/ambiguous, which is difficult to infer.

[j]EMPTY: The tag "empty (EMPTY)" is used when there is no practical meaning to determine the patient's condition, such as interpretation of some medical knowledge by the doctor, independent of the patient's current condition, inspection items, drug names, etc.

Discussion

From the scoring results, the ERNIE model, which has been trained on a large Chinese corpus, had a total score 12.6 points higher than that of the BERT model in our task. BERT-WWM surpassed ERNIE, with a score of 69.28. Our RoBERTa-WWM-ext + CNN model improved the overall score by 1.28. With the addition of the message sender in the corpus of RoBERTa-WWM-ext, the correct rate of answering sentences also improved.

A previous report assessed BERT fine-tuning as embedding input into the text CNN model and showed that the accuracy rate was 0.31% higher than that of the original BERT model and was more stable [28]. We used CNN to compute sentence features. To verify our model's knowledge learning ability, we chose ERNIE [29], original BERT, and Chinese BERT with WWM to do the same task, and then compared the results of these models.

In this study, we showed that our model outperformed the other models on the task. The test was not manually modified, and the error of the training data limited the role of manual rules. We tried to add rules to correct the positive labeling, but the total score was only 29.31 points. The accuracy of the positive label was 92.33, but the recall was only 16.46. Due to the false-positive interference of the original data, it was difficult to improve the accuracy of the model itself through artificial rules. The longest sequence length supported by BERT is 512. The text tasks suitable for processing include short texts, such as comments on social platforms and article titles, but for a medical dialogue composed of more than 50 single sentences, the length is obviously not enough. We can only use the truncation method to preprocess text, that is, first truncation, tail truncation, and head to tail truncation, which adds some difficulty to the preliminary work. According to the work of Zeng et al, the base model did improve the accuracy rate by adjusting the downstream tasks [30]. For the single model, XLNET and RoBERTa were better than BERT and ERNIE,

and the integration of multiple models will improve the model by 2.58% on average. The results of this study indicated that the accuracy of the model improved with small and middle sample sizes. The multimodel joint integration was an effective way to improve the accuracy of the entity attribute annotation.

"Internet medical+" was part of China's rapid development after "Internet+" became China's national strategy in 2015 [31]. In 2019, the novel coronavirus pneumonia outbreak occurred globally, and traditional medical treatment brought many malpractices, which stimulated the technical development of internet inquiry [32]. In the 9th IEEE International Conference on Health Care Informatics (ICHI) in 2021, some scholars proposed to integrate structured data with unstructured text annotation recorded in the classification stage, and use NLP methods for admission prediction and triage notes [33]. This study hopes to further optimize medical information and pave the way for the automatic generation of medical cases through the automatic entity annotation of doctor-patient real dialogue text generated in the process of consultation. It is speculated that our study findings will contribute to the application of NLP methods in the field of health care.

Acknowledgments

The authors would like to thank the 7th China Health Information Processing Conference organizers for providing the training, development, and test corpora. This research is funded by the National Key Research and Development Program of China (grant ID: 2020AAA0104905).

Conflicts of Interest

None declared.

References

1. Jiang X, Xie H, Tang R, Du Y, Li T, Gao J, et al. Characteristics of Online Health Care Services From China's Largest Online Medical Platform: Cross-sectional Survey Study. J Med Internet Res 2021 Apr 15;23(4):e25817 [FREE Full text] [doi: 10.2196/25817] [Medline: 33729985]
2. 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]
3. Dang Y, Guo S, Guo X, Vogel D. Privacy Protection in Online Health Communities: Natural Experimental Empirical Study. J Med Internet Res 2020 May 21;22(5):e16246 [FREE Full text] [doi: 10.2196/16246] [Medline: 32436851]
4. Yaghoobzadeh Y, Adel H, Schuetze H. Corpus-Level Fine-Grained Entity Typing. Journal of Artificial Intelligence 2018 Apr 17;61:835-862. [doi: 10.1613/jair.5601]
5. Xu B, Luo Z, Huang L, Liang B, Xiao Y, Yang D, et al. METIC: Multi-Instance Entity Typing from Corpus. In: CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018 Presented at: 27th ACM International Conference on Information and Knowledge Management; October 22-26, 2018; Torino, Italy p. 903-912. [doi: 10.1145/3269206.3271804]
6. Wentong L, Yanhui Z, Fei Z, Xiangbing J. Named Entity Recognition of Electronic Medical Records Based on BERT. Journal of Hunan University of Technology 2020;34(4):34-62. [doi: 10.3969/j.issn.1673-9833.2020.04.009]
7. Piao Y, Wenyong D. Chinese Named Entity Recognition Method Based on BERT Embedding. Computer Engineering 2020;46(4):52-55. [doi: 10.19678/j.issn.1000-3428.0054272]
8. Dun-Wei G, Yong-Kai Z, Yi-Nan G, Bin W, Kuan-Lu F, Yan H. Named entity recognition of Chinese electronic medical records based on multifeature embedding and attention mechanism. Chinese Journal of Engineering 2021;43(9):1190-1196. [doi: 10.13374/j.issn2095-9389.2021.01.12.006]
9. Xue D, Zhipeng J, Yi G. Cross-department chunking based on Chinese electronic medical record. Application Research of Computers 2017;34(7):2084-2087 [FREE Full text]
10. Zhang Y, Wang X, Hou Z, Li J. Clinical Named Entity Recognition From Chinese Electronic Health Records via Machine Learning Methods. JMIR Med Inform 2018 Dec 17;6(4):e50 [FREE Full text] [doi: 10.2196/medinform.9965] [Medline: 30559093]
11. Jiang J, Cameron A, Yang M. Analysis of Massive Online Medical Consultation Service Data to Understand Physicians' Economic Return: Observational Data Mining Study. JMIR Med Inform 2020 Feb 18;8(2):e16765 [FREE Full text] [doi: 10.2196/16765] [Medline: 32069213]
12. Che W, Zhao Y, Guo H, Su Z, Liu T. Sentence Compression for Aspect-Based Sentiment Analysis. IEEE/ACM Trans. Audio Speech Lang. Process 2015 Dec;23(12):2111-2124. [doi: 10.1109/taslp.2015.2443982]
13. Wang T, Lu K, Chow KP, Zhu Q. COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. IEEE Access 2020;8:138162-138169. [doi: 10.1109/access.2020.3012595]
14. Kummervold PE, Martin S, Dada S, Kilich E, Denny C, Paterson P, et al. Categorizing Vaccine Confidence With a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse. JMIR Med Inform 2021 Oct 08;9(10):e29584 [FREE Full text] [doi: 10.2196/29584] [Medline: 34623312]

15. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. Alibaba Group. 2021. URL: https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414 [accessed 2022-03-22]
16. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics 1996 Sep;5(3):299-314. [doi: 10.1080/10618600.1996.10474713]
17. Pattanayak S. Convolutional Neural Networks. In: Pro Deep Learning with TensorFlow. Berkeley, CA: Apress; 2017:153-221.
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. 2019. URL: https://arxiv.org/abs/1907.11692 [accessed 2022-03-22]
19. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-Training With Whole Word Masking for Chinese BERT. IEEE/ACM Trans. Audio Speech Lang. Process 2021;29:3504-3514. [doi: 10.1109/taslp.2021.3124365]
20. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv. 2020. URL: https://arxiv.org/abs/1910.03771 [accessed 2022-03-22]
21. Lei T, Barzilay R, Jaakkola T. Molding CNNs for text: non-linear, non-consecutive convolutions. arXiv. 2015. URL: https://arxiv.org/abs/1508.04112 [accessed 2022-03-22]
22. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2022-03-22]
23. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention Is All You Need. arXiv. 2017. URL: https://arxiv.org/abs/1706.03762 [accessed 2022-03-22]
24. Dandan D, Jiashan T, Yong W, Kehai Y, Jiashan T, Yong W. Chinese Short Text Classification Algorithm Based on BERT Model. Computer Engineering 2021;47(1):47-86. [doi: 10.19678/j.issn.1000-3428.0056222]
25. Dauphin Y, Fan A, Auli M, Grangier D. Language Modeling with Gated Convolutional Networks. arXiv. 2017. URL: https://arxiv.org/abs/1612.08083 [accessed 2022-03-22]
26. Chen Z, Qian T. Transfer Capsule Network for Aspect Level Sentiment Classification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy p. 547-556. [doi: 10.18653/v1/p19-1052]
27. Kun W, Yi Z, Shuya F, Shouyin L. Long text aspect-level sentiment analysis based on text filtering and improved BERT. Journal of Computer Applications 2020;40(10):2838-2844. [doi: 10.11772/j.issn.1001-9081.2020020164]
28. Xiaowei Z, Jianfei S. Research on News Text Classification Based on Improved BERT-CNN Model. Video Engineering 2021;45(7):146-150. [doi: 10.16280/j.videoe.2021.07.040]
29. Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, et al. ERNIE: Enhanced Representation through Knowledge Integration. arXiv. 2019. URL: https://arxiv.org/abs/1904.09223 [accessed 2022-03-22]
30. Zeng K, Xu Y, Lin G, Liang L, Hao T. Automated classification of clinical trial eligibility criteria text based on ensemble learning and metric learning. BMC Med Inform Decis Mak 2021 Jul 30;21(Suppl 2):129 [FREE Full text] [doi: 10.1186/s12911-021-01492-z] [Medline: 34330259]
31. Xiaoyan Z, Jing B, Rong L. Studying on The Existing Modes of “Internet Plus Medical Services” in China. Chinese Health Service Management 2019;36(1):8 [FREE Full text]
32. Hui C, Qiong Z, Xiaoli L, Bochun Y. Opportunity and Reflection of the lnternet+Medical Under COVlD-19 Epidemic Situation. Chinese Hospital Management 2020;40(6):38-40 [FREE Full text]
33. Arnaud É, Elbattah M, Gignon M, Dequen G. NLP-Based Prediction of Medical Specialties at Hospital Admission Using Triage Notes. 2021 Presented at: 9th International Conference on Healthcare Informatics (ICHI); August 9-12, 2021; Victoria, BC, Canada p. 548-553. [doi: 10.1109/ICHI52183.2021.00103]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CNN: convolutional neural network
ERNIE: Enhanced Representation through Knowledge Integration
MLM: masked language model
NLP: natural language processing
NSP: next sentence prediction
RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach
RoBERTa-WWM-ext: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach with whole word masking extended
WWM: whole word masking

Edited by C Lovis; submitted 10.12.21; peer-reviewed by M Elbattah, H Monday; comments to author 13.02.22; revised version received 24.02.22; accepted 25.02.22; published 21.04.22.

Please cite as:
Sun Y, Gao D, Shen X, Li M, Nan J, Zhang W
Multi-Label Classification in Patient-Doctor Dialogues With the RoBERTa-WWM-ext + CNN (Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach With Whole Word Masking Extended Combining a Convolutional Neural Network) Model: Named Entity Study
JMIR Med Inform 2022;10(4):e35606
URL: https://medinform.jmir.org/2022/4/e35606
doi:10.2196/35606
PMID:35451969

Original Paper

Using Natural Language Processing and Machine Learning to Preoperatively Predict Lymph Node Metastasis for Non–Small Cell Lung Cancer With Electronic Medical Records: Development and Validation Study

Danqing Hu[1*], MSc; Shaolei Li[2*], MD; Huanyao Zhang[1], MSc; Nan Wu[2], MD; Xudong Lu[1], PhD

[1]College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou, China

[2]Department of Thoracic Surgery II, Peking University Cancer Hospital and Institute, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Xudong Lu, PhD
College of Biomedical Engineering and Instrumental Science
Zhejiang University
38 Zheda Road
Hangzhou, 310027
China
Phone: 86 139 5711 8891
Email: lvxd@zju.edu.cn

Abstract

Background: Lymph node metastasis (LNM) is critical for treatment decision making of patients with resectable non–small cell lung cancer, but it is difficult to precisely diagnose preoperatively. Electronic medical records (EMRs) contain a large volume of valuable information about LNM, but some key information is recorded in free text, which hinders its secondary use.

Objective: This study aims to develop LNM prediction models based on EMRs using natural language processing (NLP) and machine learning algorithms.

Methods: We developed a multiturn question answering NLP model to extract features about the primary tumor and lymph nodes from computed tomography (CT) reports. We then combined these features with other structured clinical characteristics to develop LNM prediction models using machine learning algorithms. We conducted extensive experiments to explore the effectiveness of the predictive models and compared them with size criteria based on CT image findings (the maximum short axis diameter of lymph node >10 mm was regarded as a metastatic node) and clinician's evaluation. Since the NLP model may extract features with mistakes, we also calculated the concordance correlation between the predicted probabilities of models using NLP-extracted features and gold standard features to explore the influence of NLP-driven automatic extraction.

Results: Experimental results show that the random forest models achieved the best performances with 0.792 area under the receiver operating characteristic curve (AUC) value and 0.456 average precision (AP) value for pN2 LNM prediction and 0.768 AUC value and 0.524 AP value for pN1&N2 LNM prediction. And all machine learning models outperformed the size criteria and clinician's evaluation. The concordance correlation between the random forest models using NLP-extracted features and gold standard features is 0.950 and improved to 0.984 when the top 5 important NLP-extracted features were replaced with gold standard features.

Conclusions: The LNM models developed can achieve competitive performance using only limited EMR data such as CT reports and tumor markers in comparison with the clinician's evaluation. The multiturn question answering NLP model can extract features effectively to support the development of LNM prediction models, which may facilitate the clinical application of predictive models.

(JMIR Med Inform 2022;10(4):e35475) doi:10.2196/35475

KEYWORDS

non–small cell lung cancer; lymph node metastasis prediction; natural language processing; electronic medical records; lung cancer; prediction models; decision making; machine learning; algorithm; forest modeling

Introduction

Lung cancer remains the leading cause of cancer death worldwide, representing approximately 1 in 5 (18.0%) cancer deaths [1]. Non–small cell lung cancer (NSCLC) accounts for about 84% of lung cancer cases, and its 5-year relative survival rate is only 25.0% [2], making it one of the biggest threats to human health.

Staging of NSCLC is a process to determine the extent of the cancer and is critical to prognosis evaluation and treatment decision making [3,4]. The TNM stage classification [5] is the most widely used staging method in clinical practice; it describes the anatomic extent of a tumor from 3 aspects (ie, T for extent of the primary tumor, N for involvement of lymph nodes, M for distant metastases). For patients with resectable NSCLC, preoperative confirmed N2 (a type of N stage) lymph node metastasis (LNM) indicates neoadjuvant therapy should be given before surgery to achieve the best clinical practice [3]. Currently, various advanced noninvasive diagnostic modalities are available for N staging like chest computed tomography (CT) and positron emission tomography–computed tomography (PET-CT). In clinical practice, clinicians commonly use a size criterion (ie, the maximum short axis diameter of lymph node >10 mm on CT scan) to discriminate LNM from benign nodes and yield 55% sensitivity [6]. Another criterion is the maximum standardized uptake value (SUVmax) of lymph node >2.5 on PET-CT scan, which has an 81% sensitivity [7]. Invasive methods such as mediastinoscopy and endobronchial ultrasound-guided transbronchial needle aspiration have better diagnostic abilities than noninvasive methods. However, these methods are mainly for lymph nodes with indications and not suitable for patients with severe comorbidities, so they are not routinely used in clinical practice [8]. One study analyzed data from 9 clinical trials and found nearly 38% of patients were misclassified in comparison with their pathological N staging [9]. Therefore, new reliable LNM prediction methods are required to alleviate this clinical dilemma.

For precise staging, researchers explored using statistical analysis or machine learning methods to learn nontrivial knowledge between the comprehensive patient features and LNM status [8,10-16]. Recently, with the rapid development of hospital information systems, a large volume of electronic medical records (EMR) has become available, and it contains almost all clinical features about patients. However, some important features are recorded in the narratives in free text, such as the size of the tumor and lymph node, tumor density, pleural indentation, etc, which hinders their direct use. Manual extraction is time-consuming and error-prone. So, one big challenge is how to extract this information effectively to support subsequent tasks like LNM prediction [17]. A review by Garg et al [18] found studies in which users were automatically prompted to use the system achieved better performance in comparison with those in which users were required to actively initiate the system. The finding implicitly indicates that the duplicative data entry activity may explain why the predictive models are not widely adopted in the clinic despite their potential to improve diagnostic accuracy. Furthermore, with the prevalence of machine learning models, more features are required for analysis, making the clinical application of the models more difficult [19-21].

Natural language processing (NLP) offers the opportunity to automatically extract information to support the application of predictive models [17,22]. Many studies used rule-based, machine learning, or deep learning methods to extract the cancer-related information from free-text EMR data [22-29], but only a few included further elaboration on how to exploit the extracted information. Chen et al [30] extracted information from various clinical notes including CT reports and operative notes to calculate the Cancer of the Liver Italian Program score. Martinez et al [31] extracted information from pathology reports to calculate the TNM and Australian clinicopathological stage of colorectal cancer. Castro et al [32] developed an NLP system for automated breast imaging reporting and data system (BI-RADS) categories extraction from breast radiology reports. Bozkurt et al [33,34] developed an information extraction pipeline to extract information from mammography reports to predict the malignancy of breast cancer. Sui et al [35] constructed an NLP-based feature generalizing to extract features from free-text EMR data and provided the stage of lung cancer using a Bayesian reasoning network. Yuan et al [36] used NLP tools to extract multiple features from EMRs to estimate survival for patients with lung cancer. Although many studies have explored how to extract the cancer-related information from various types of free-text narratives and some also exploit the extracted information for cancer risk evaluation, diagnosis, and pathological staging, few studies exploit the extracted information from radiological reports for preoperative LNM prediction, especially for NSCLC.

In this study, we aim to use EMR data to develop LNM prediction models for NSCLC patients. We first developed a multiturn question answering NLP model to extract the features from CT reports and then combined these features with other clinical characteristics to develop the predictive models. Since the NLP model may produce imperfect extraction results, we also conducted experiments to compare the predicted probabilities between models using NLP-extracted features and gold standard features.

Methods

Patients

We retrospectively analyzed EMR data of 794 patients who underwent surgical resection for NSCLC with systematic mediastinal lymphadenectomy at the Department of Thoracic Surgery II of Peking University Cancer Hospital from 2010 to 2018. All patients underwent contrast-enhanced chest CT images within 2 months before surgical resection. We excluded the patients with preoperative chemotherapy or radiotherapy. The collected EMR includes demographic information, medical history, CT reports, preoperative serum tumor markers, and pathology reports, which can be analyzed to develop the prediction model. For each patient, we also collected the clinical staging that clinicians evaluated before surgery as the baseline to compare with the LNM prediction models.

Ethics Approval

This study was approved by the Ethics Committee of Peking University Cancer Hospital (2019KT59).

Clinical and Pathological LNM Evaluation

In this study, all included patients underwent systematic mediastinal lymphadenectomy during surgical resection. The lymph node tissues were examined by pathologists, and the metastasis results were recorded in the postoperative pathology reports. We reviewed the pathology reports to determine the LNM status and label the pathological N (pN) stage (pN0/pN1/pN2) for each patient based on the 8th edition TNM stage classification [5] as the gold standard. We also used the size criterion (ie, the maximum short axis diameter of lymph node >10 mm on CT scan as positive) to label the clinical N (cN) stages (cN0/cN1/cN2) based on the CT-reported lymph node size. Moreover, we collected the cN stages, which were determined preoperatively by a thoracic surgeon using all available patient data including the information used in this study. The thoracic surgeon has 10 years of experience in lung cancer surgery. The cN stages determined by the size criterion and the thoracic surgeon were regarded as the baselines.

NLP Feature Extraction

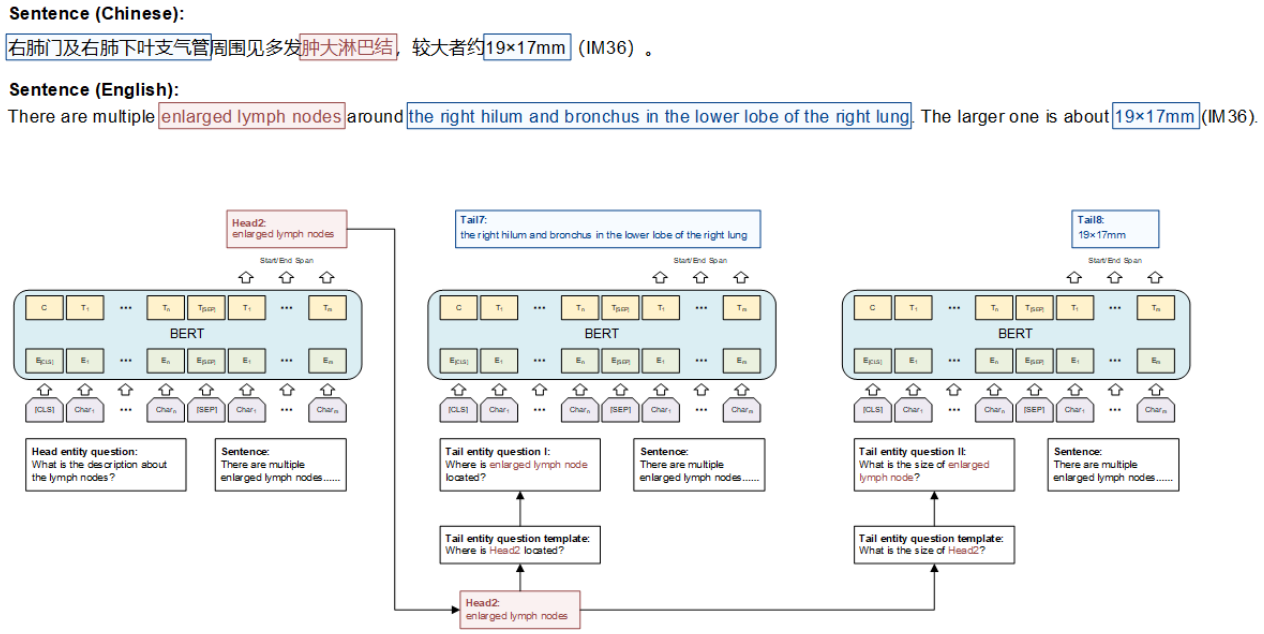
As one of the most important preoperative examinations, CT reports record valuable information about the tumors and lymph nodes, which is of paramount importance for staging. However, the free-text nature of CT reports makes it difficult to understand and analyze them using computer programs. In our previous work [27], we developed an information extraction system composed of named entity recognition, relation classification, and postprocessing modules to extract valuable information in a pipeline manner. However, in this pipeline, the subsequent tasks would be influenced by the outputs of former tasks, which may affect the performance of the whole system. Therefore, to alleviate this problem, we applied a multiturn question answering (MTQA) [37] approach to extract information from CT reports in this study. Using the MTQA strategy, we can encode the relation into the question query and jointly model entity and relation in a natural question answering way.

Specifically, we first defined 10 questions related to the primary tumor and lymph nodes. All questions are listed in Table 1. Note that there are 2 types of questions (ie, head entity questions and tail entity question templates). In the model training stage, we inserted the annotated head entities into the slots in the tail entity question templates as the tail entity questions. We then used 2 special tokens (ie, CLS and SEP) to concatenate the questions and sentences in the reports as the inputs and annotated entities as the answers to conduct the bidirectional encoder representations from transformers (BERT) model training. In the model test stage, we first concatenated the head entity questions and sentences in the reports as the inputs and applied the trained MTQA model to extract the head entities (ie, tumor and lymph node). If there were any head entities recognized, we inserted the extracted head entities into the slots in the tail entity question templates as the tail entity questions and combined them with sentences in the reports as the inputs to drive the tail entity extraction. A case of the MTQA application is shown in Figure 1. Finally, the extracted head and tail entities are organized as triples, and a rule-based postprocessing algorithm proposed in the previous work [27] is used to process the triples to obtain the standardized NLP-extracted features. Furthermore, the NLP-extracted features were manually reviewed and corrected by a clinician based on the report contents as the gold standard features. In this study, we used BERT [38], an advanced pretrained language representation model, to tag the answer for each question.

Table 1. Questions and entity types for natural language processing–extracted features.

Question (Chinese)	Question (English)	Answer notation	Entity type
Head entity question			
原发肿物的相关描述是什么？	What is the description about the primary tumor?	Head1	Tumor
淋巴结的相关描述是什么？	What is the description about the lymph nodes?	Head2	Lymph node
Tail entity question template			
Head1 位于什么地方？	Where is Head1 located?	Tail1	Location
Head1 的大小是多少？	What is the size of Head1?	Tail2	Size
Head1 的形状是什么？	What is the shape of Head1?	Tail3	Shape
Head1 的密度是什么？	What is the density of Head1?	Tail4	Density
与Head1 相关的胸膜侵犯的描述是什么？	What is the description about the pleura invasion related to Head1?	Tail5	Pleura
与Head1 相关的血管侵犯的描述是什么？	What is the description about the vessel invasion related to Head1?	Tail6	Vessel
Head2 位于什么地方？	Where is Head2 located?	Tail7	Location
Head2 的大小是多少？	What is size of Head2?	Tail8	Size

Figure 1. A case of multiturn question answering application. BERT: bidirectional encoder representations from transformers.



LNM Prediction

Six machine learning algorithms were applied to develop the LNM prediction models, including logistic regression (LR) [39], L2-logistic regression (L2-LR) [40], random forest (RF) [41], LightGBM (LGBM) [42], support vector machine (SVM) [43], and artificial neural network (ANN) [44]. LR is the conventional classification method, and L2-LR is the LR with the L2 regularization for parameters. RF and LGBM are ensemble methods but with different ways to combine the weak decision trees. SVM is a classical algorithm that constructs hyperplanes in a high- or infinite-dimensional space to classify samples. ANN is a supervised learning algorithm that can learn nonlinear functions between features and targets. LR and L2-LR have good interpretability because the predicted results can be calculated by a simple linear function and a sigmoid transformation. RF and LGBM are also interpretable, in which they can provide the feature importance.

Experimental Setup

In this study, we used the Whole Word Masking version of BERT [45] pretrained on the Chinese Wikipedia corpus as the tagging model in the MTQA. An additional 359 annotated CT reports from our previous work were used to develop and evaluate the MTQA model. We randomly split 70% of CT reports as the training set, 10% as the validation set, and 20% as the test set. A total of 100 of these reports were each annotated by 2 biomedical informatics engineers to calculate the interannotator agreement score using the kappa score. Pipeline methods with bidirectional long short-term memory (BiLSTM) and BERT were selected as the baseline. To obtain the NLP-extracted features for LNM prediction, the MTQA model developed on the 359 reports was used to process the 794 CT reports of included patients. Subsequently, the NLP-extracted features were manually reviewed and corrected by a clinician as the gold standard features.

Univariate analysis was performed using the Mann-Whitney *U* test for continuous features and Pearson chi-square test for categorical features. $P<.05$ was considered statistically significant. To obtain robust experimental results, a 10-fold cross-validation strategy was first performed on the total data set. The 10-fold cross-validation randomly split the data set into 10 subsets. Each subset was considered as the independent test set and the remaining 9 subsets were considered as the training set. During each fold, a 5-fold cross-validation was applied on the training set to find the optimal hyperparameters for the machine learning algorithms by a grid search. When the optimal hyperparameters were selected, we retrained the prediction model on the training set and tested it on the test set to obtain the final predictive performance. Using this strategy, we can ensure that the test set is always invisible during the model training and hyperparameter tuning and obtain the predicted probability for each case. The hyperparameter spaces are as follows:

- LR: tol ∈ {1e–3, 1e–4, 1e–5}, max_iter ∈ {500, 1000}
- L2-LR: C ∈ {10, 1, 0.1}, tol ∈ {1e–3, 1e–4, 1e–5}, max_iter ∈ {500, 1000}
- RF: n_estimators ∈ {50, 100, 200}, max_depth ∈ {2, 3}, min_samples_leaf ∈ {1, 2}
- LGBM: n_estimators ∈ {50, 100, 200}, max_depth ∈ {2, 3}, num_leaves ∈ {20, 31, 50}, min_child_samples ∈ {1, 2, 3}, reg_alpha ∈ {2, 3}
- SVM: C ∈ {10, 1, 0.1, 0.01}, kernel ∈ {‘linear,’ ‘rbf,’ ‘poly’}, tol ∈ {1e–3, 1e–4, 1e–5}
- ANN: hidden_layer_sizes ∈ {5, 10, 30}, learning_rate ∈ {1e–2, 1e–3, 1e–4}, alpha ∈ {1e–3, 1e–4, 1e–5}

We applied the receiver operating characteristic (ROC) curve to evaluate the diagnostic performances of the machine learning models. Besides the ROC curve, we also used the precision-recall (PR) curve to test the models because the ROC curve pays attention to sensitivity and specificity but ignores precision. The mean area under the receiver operating characteristic curve (AUC) and average precision (AP) values

with standard derivations were calculated based on the 10-fold cross-validation results. We also drew the ROC curves and PR curves to compare with the size criterion (maximum short axis diameter of lymph node >10 mm on CT) and the clinician's evaluation. All LNM prediction models were developed using the Scikit-learn 0.24.1 and LightGBM 3.2.0 Python packages. All statistical analyses were conducted using SciPy 1.6.2 Python package.

Results

Patient Characteristics

Table 2 shows the characteristics of all 794 patients. Univariate analysis was performed for all collected features, and 13.2% (105/794) of patients had pN2 LNM. Sex, age, drinking history, family history, and disease history are not significantly associated with the pN2. The pN2 occurred more frequently in smokers (P=.04). The long and short axis diameters of the tumor in pN2 patients are significantly larger than those in pN0 and pN1 patients (both P<.001). Patients with solid nodules are more likely to have pN2 (P<.001). Other morphological characteristics of tumor-like lobulation and pleural indentation are more likely to occur in pN2 patients (P=.006 and P=.003, respectively), but spiculation and vessel invasion present no significant differences between pN2 and other patients. Using 10 mm as the size criterion, the maximum long and short axis diameters of the hilar and mediastinal lymph nodes show significant differences between the 2 groups (P=.008, P<.001, P<.001, and P<.001, respectively). Among all 6 serum tumor biomarkers, carcinoembryonic antigen (CEA), carbohydrate antigen 12-5 (CA125), and neuron-specific enolase (NSE) show significant differences between the 2 groups (P<.001, P<.001, and P=.048, respectively).

Table 2. Patient characteristics.

	Total (n=794)	LNM[a] status		*P* value
		pN2[b] (n=105)	pN0[c] or pN1[d] (n=689)	
Age (years), mean (SD)	60.92 (51.48 to 70.36)	60.87 (51.87 to 69.86)	60.93 (51.42 to 70.44)	.45
Sex, n (%)	—[e]	—	—	.06
Male	397	62	335	—
Female	397	43	354	—
Smoking history, n (%)	—	—	—	.04
Yes	337	55	282	—
No	457	50	407	—
Drinking history, n (%)	—	—	—	.94
Yes	183	25	158	—
No	611	80	531	—
Family history, n (%)	—	—	—	.32
Yes	137	14	123	—
No	657	91	566	—
Hypertension, n (%)	—	—	—	.18
Yes	232	37	195	—
No	562	68	494	—
Diabetes, n (%)	—	—	—	.25
Yes	84	15	69	—
No	710	90	620	—
Pulmonary tuberculosis, n (%)	—	—	—	.33
Yes	33	2	31	—
No	761	103	658	—
Cardiovascular disease, n (%)	—	—	—	.06
Yes	36	9	27	—
No	758	96	662	—
Cerebrovascular disease, n (%)	—	—	—	.35
Yes	29	6	23	—
No	765	99	666	—
Tumor location[f], n (%)	—	—	—	.22
RUL[g]	249	27	222	—
RML[h]	59	4	55	—
RLL[i]	150	18	132	—
LUL[j]	185	31	154	—
LLL[k]	126	21	105	—
Other	25	4	21	—
TLA[f,l], median (IQR)	2.61 (1.20 to 4.01)	3.02 (1.64 to 4.39)	2.55 (1.15 to 3.94)	<.001
TSA[f,m], median (IQR)	2.03 (0.88 to 3.18)	2.38 (1.27 to 3.48)	1.98 (0.83 to 3.13)	<.001
Spiculation[f], n (%)	—	—	—	.08
Yes	255	42	213	—

	Total (n=794)	LNM[a] status		*P* value
		pN2[b] (n=105)	pN0[c] or pN1[d] (n=689)	
No	539	63	476	—
Lobulation[f], n (%)	—	—	—	<.001
Yes	211	48	163	—
No	583	57	526	—
Tumor density[f], n (%)	—	—	—	<.001
pGGO[n]	124	0	124	—
mGGO[o]	96	3	93	—
Solid nodule	574	102	472	—
Vessel invasion[f], n (%)	—	—	—	.87
Yes	52	6	46	—
No	742	99	643	—
Pleural indentation[f], n (%)	—	—	—	.001
Yes	406	70	336	—
No	388	35	353	—
HLNLA[f,p], n (%)	—	—	—	.008
>10 mm	148	30	118	—
≤10 mm	646	75	571	—
HLNSA[f,q], n (%)	—	—	—	<.001
>10 mm	66	19	47	—
≤10 mm	728	86	642	—
MLNLA[f,r], n (%)	—	—	—	<.001
>10 mm	191	50	141	—
≤10 mm	603	55	548	—
MLNSA[f,s], n (%)	—	—	—	<.001
>10 mm	72	27	45	—
≤10 mm	722	78	644	—
CEA[t], median (IQR)	5.31 (–6.66 to 17.27)	12.66 (–8.44 to 33.76)	4.18 (–5.17 to 13.54)	<.001
CA199[u], median (IQR)	14.41 (–3.24 to 32.06)	15.80 (–5.08 to 36.68)	14.20 (–2.90 to 31.29)	.47
CA125[v], median (IQR)	14.46 (0.03 to 28.90)	19.88 (–5.56 to 45.32)	13.64 (1.96 to 25.32)	<.001
NSE[w], median (IQR)	15.81 (8.85 to 22.78)	16.26 (10.19 to 22.33)	15.75 (8.66 to 22.83)	.048
Cyfra211[x], median (IQR)	3.20 (–0.23 to 6.62)	3.55 (–0.64 to 7.75)	3.14 (–0.15 to 6.43)	.06

	Total (n=794)	LNM[a] status		*P* value
		pN2[b] (n=105)	pN0[c] or pN1[d] (n=689)	
SCCAg[y], median (IQR)	0.96 (–0.16 to 2.08)	1.18 (–0.62 to 2.99)	0.93 (–0.04 to 1.90)	.14

[a]LNM: lymph node metastasis.
[b]pN2: pathological N stage 2.
[c]pN0: pathological N stage 0.
[d]pN1: pathological N stage 1.
[e]Not applicable.
[f]Features recorded in computed tomography reports.
[g]RUL: right upper lobe.
[h]RML: right middle lobe.
[i]RLL: right lower lobe.
[j]LUL: left upper lobe.
[k]LLL: left lower lobe.
[l]TLA: tumor long axis.
[m]TSA: tumor short axis
[n]pGGO: pure ground glass opacity.
[o]mGGO: mixed ground glass opacity.
[p]HLNLA: hilar lymph node long axis.
[q]HLNSA: hilar lymph node short axis.
[r]MLNLA: mediastinal lymph node long axis.
[s]MLNSA: mediastinal lymph node short axis.
[t]CEA: carcinoembryonic antigen.
[u]CA199: carbohydrate antigen 19-9.
[v]CA125: carbohydrate antigen 12-5.
[w]NSA: neuron-specific enolase.
[x]Cyfra211: cytokeratin 19-fragments.
[y]SCCAg: squamous cell carcinoma antigen.

Performance of pN2 LNM Prediction Models

As preoperative confirmed N2 indicating neoadjuvant therapy should be given before surgery, we first developed machine learning models to predict the pN2 LNM. We regarded the pN2 patients as positive and pN0 and pN1 patients as negative to train the predictive models. To obtain reliable models, we used the gold standard features instead of NLP-extracted features in this section. Table 3 shows the performances of all models. The RF model achieved the highest averaged AUC value with 0.792 and the LGBM model achieved the highest averaged AP value with 0.457 while all models' 95% CI are overlapping with each other. The LR obtained a competitive performance in comparison with ANN and SVM. The L2-LR did not obtain improvements in AUC value and AP value compared with the LR. To compare with the size criterion and clinician's evaluation, we used the probabilities predicted during the 10-fold cross-validation to draw the ROC and PR curves. Figure 2 shows the ROC curves and PR curves of pN2 prediction models and the results of the size criterion and clinician's evaluation. From Figure 2 we can notice all the ROC curves and PR curves are above the points of size criterion and clinician's evaluation, which indicates the developed pN2 prediction models not only have better discriminative ability than the diagnostic size criterion used in the clinical practice but also may exceed the clinician in pN2 LNM evaluation.

Table 3. Performances of pN2 lymph node metastasis prediction models.

Model	AUC[a]			AP[b]		
	Mean	SD	95% CI	Mean	SD	95% CI
LR[c]	0.778	0.041	0.747-0.809	0.442	0.075	0.385-0.499
L2-LR[d]	0.768	0.038	0.739-0.796	0.413	0.072	0.359-0.467
ANN[e]	0.769	0.051	0.730-0.808	0.434	0.095	0.363-0.506
SVM[f]	0.771	0.071	0.718-0.825	0.453	0.084	0.389-0.516
RF[g]	0.792	0.042	0.760-0.825	0.456	0.075	0.399-0.512
LGBM[h]	0.787	0.044	0.755-0.820	0.457	0.101	0.381-0.534

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

[b]AP: average precision.

[c]LR: logistic regression.

[d]L2-LR: L2-logistic regression.

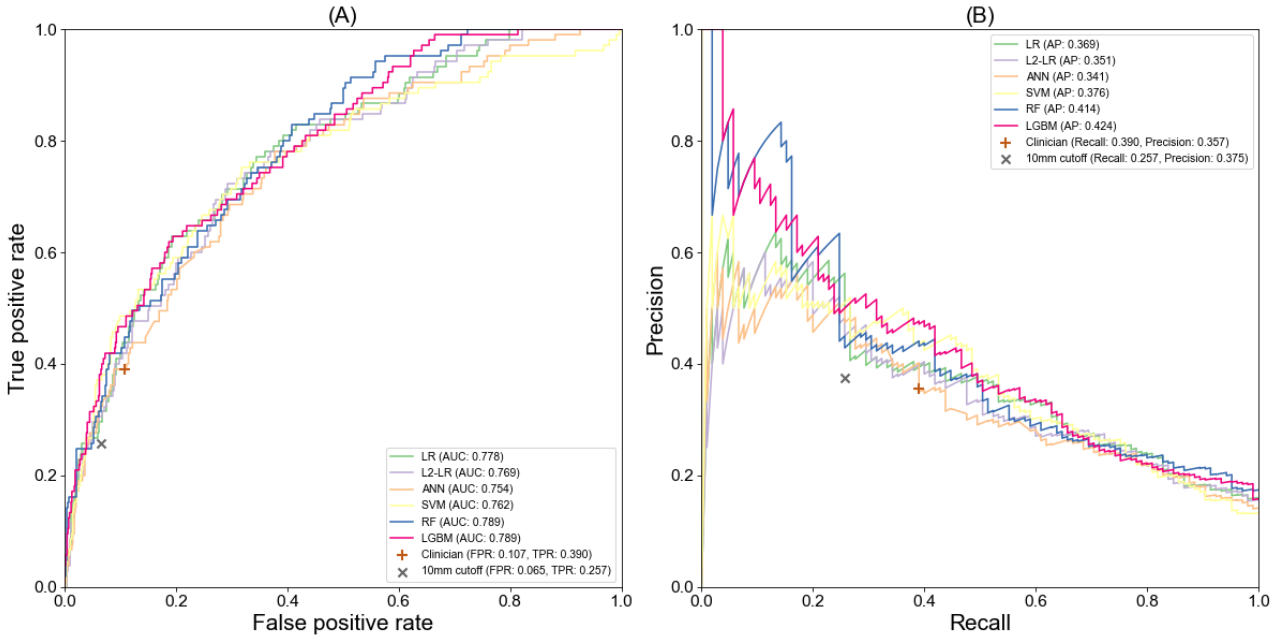
[e]ANN: artificial neural network.

[f]SVM: support vector machine.

[g]RF: random forest.

[h]LGBM: LightGBM.

Figure 2. The receiver operating characteristic curve (A) and precision-recall curves (B) of pN2 prediction models.



Performance of pN1&N2 LNM Prediction Models

Besides predicting pN2 LNM, we also developed machine learning models to predict the pN1&N2 LNM by regarding patients with pN1 or pN2 LNM as positive. The model training and evaluation processes are the same as pN2 LNM prediction. Table 4 shows the performances of the machine learning models for pN1&N2 LNM prediction. LGBM obtained the highest averaged AUC value with 0.771. The RF model achieved a comparable performance in comparison with LGBM. As in pN2 prediction, LGBM and RF obtained better predictive performances than other models. Figure 3 shows the ROC curves and PR curves of pN1&N2 LNM prediction models. The curves of the machine learning models are also all above the points of the size criterion and clinician's evaluation.

Table 4. Performances of pN1&N2 lymph node metastasis prediction models.

Model	AUC[a]			AP[b]		
	Mean	SD	95% CI	Mean	SD	95% CI
LR[c]	0.740	0.035	0.714-0.766	0.467	0.058	0.423-0.510
L2-LR[d]	0.736	0.044	0.704-0.769	0.465	0.058	0.422-0.509
ANN[e]	0.734	0.047	0.698-0.770	0.479	0.087	0.413-0.545
SVM[f]	0.735	0.023	0.717-0.752	0.474	0.047	0.439-0.509
LGBM[g]	0.768	0.030	0.745-0.791	0.524	0.044	0.491-0.557
RF[h]	0.771	0.026	0.752-0.791	0.524	0.057	0.481-0.567

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

[b]AP: average precision.

[c]LR: logistic regression.

[d]L2-LR: L2-logistic regression.

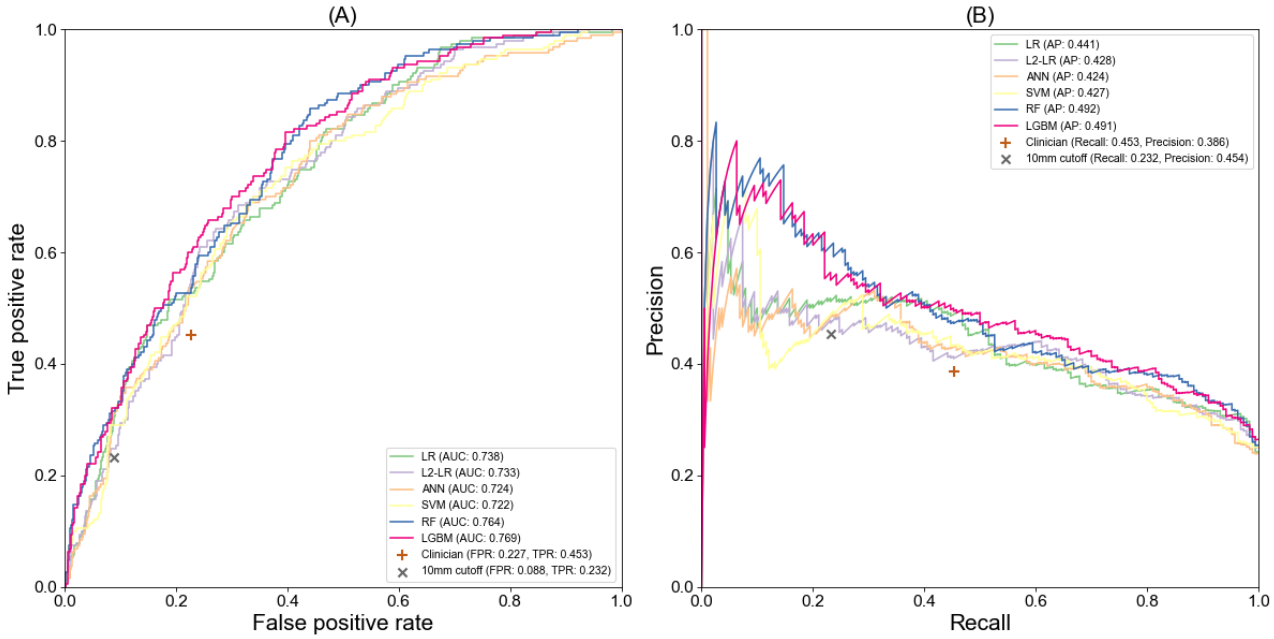
[e]ANN: artificial neural network.

[f]SVM: support vector machine.

[g]RF: random forest.

[h]LGBM: LightGBM.

Figure 3. The receiver operating characteristic curve (A) and precision-recall curves (B) of pN1&N2 prediction models.



Feature Importance

Among all machine learning models, the LR, L2-LR, RF, and LGBM can provide the feature importance. Table 5 shows the top 10 important features of LR, L2-LR, RF, and LGBM for pN2 LNM prediction. The features were ranked by averaging the weights of models developed from 10-fold cross validation. Note that the LR and L2-LR models provide weights with signs, so we used the absolute values to rank the features. Because the weight magnitudes from different models vary greatly, we used the averaged rankings of features, but not the averaged weights, to find the most important features among the 4 types of models. The CEA is ranked as the most important feature to increase the risk of pN2 LNM by all models. Features recorded in CT reports account for at least half of the top 10 important features, indicating these features are of great importance for pN2 LNM prediction.

Table 5. Top 10 important features for pN2 lymph node metastasis prediction.

Rank	LR[a]		L2-LR[b]		RF[c]		LGBM[d]		All
	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight	
1	pGGO[e,f]	–10.383	CEA[g]	3.530	CEA	0.229	CEA	46.0	CEA
2	CEA	6.010	CA125[h]	3.067	CA125	0.094	Age	23.3	Solid nodule[f]
3	CA125	4.728	pGGO[f]	–1.799	Solid nodule[f]	0.094	Solid nodule[f]	18.8	CA125
4	Solid nodule[f]	3.683	Solid nodule[f]	1.773	MLNSA[f,i]	0.073	TLA[f,j]	17.6	Age
5	TLA[f]	–2.701	Age	–1.315	MLNLA[f,k]	0.072	TSA[f,l]	15.1	MLNLA[f]
6	Age	–1.908	SCCAg[m]	0.944	TLA[f]	0.054	CA125	13.3	TLA[f]
7	SCCAg	1.763	MLNLA[f]	0.896	TSA[f]	0.048	Cyfra211[n]	12.9	pGGO[f]
8	mGGO[f,o]	1.759	Pleural indentation[f]	0.836	Cyfra211	0.038	NSE[p]	12.7	SCCAg
9	RML[f,q]	–1.729	Cardiovascular disease	0.807	SCCAg	0.037	MLNLA[f]	11.6	Lobulation[f]
10	TSA[f]	1.601	Lobulation[f]	0.725	Lobulation[f]	0.036	SCCAg	9.0	TSA[f]

[a]LR: logistic regression.
[b]L2-LR: L2-logistic regression.
[c]RF: random forest.
[d]LGBM: LightGBM.
[e]pGGO: pure ground glass opacity.
[f]Features recorded in computed tomography reports.
[g]CEA: carcinoembryonic antigen.
[h]CA125: carbohydrate antigen 12-5.
[i]MLNSA: mediastinal lymph node short axis.
[j]TLA: tumor long axis.
[k]MLNLA: mediastinal lymph node long axis.
[l]TSA: tumor short axis.
[m]SCCAg: squamous cell carcinoma antigen.
[n]Cyfra211: cytokeratin 19-fragments.
[o]mGGO: mixed ground glass opacity.
[p]NSE: neuron-specific enolase.
[q]RML: right middle lobe.

NLP-Extracted Features Versus Gold Standard Features

In this study, we applied the MTQA model to extract important features from CT reports to support the development of LNM prediction models. In this section, we first conduct experiments to explore the effectiveness of the MTQA model on feature extraction and then analyze the influence of imperfect extraction results on LNM prediction.

We used an additional 359 annotated CT reports to develop the MTQA model. The interannotator agreement score was 0.937 based on the 100 reports annotated by 2 annotators. Table 6 shows the performances of the MTQA model and the pipeline models on the test set. We can notice that the BERT-MTQA model achieved significant improvement compared with the pipeline models.

Table 7 illustrates the performance of the BERT-MTQA model on the 794 CT reports of included patients. We can notice that the accuracy values of all extracted features are higher than 0.90. The F1 scores are higher than 0.90 except for lobulation, tumor density, vessel invasion, and hilar lymph node long axis. For the NLP-extracted features ranked in the top 10 important features, the mediastinal lymph node long axis (MLNLA), tumor long axis (TLA), and tumor short axis (TSA) obtained good accuracy values and F1 scores, but the F1 scores of tumor density and lobulation are not higher than 0.90.

In this study, the MTQA model generates imperfect extractions, which may influence the subsequent application. To analyze the influence on the pN2 LNM prediction, we calculated the Pearson correlation between the predicted probabilities of models using NLP-extracted features and gold standard features. Moreover, we also replaced the NLP-extracted feature with the gold standard feature one by one according to their importance in Table 5 to explore the changes in the consistency. Figure 4

shows the concordance correlations of the pN2 LNM prediction models. The RF model obtained a high concordance correlation with 0.950 when using all NLP-extracted features in comparison with using gold standard features, and the correlation increased to 0.984 when replacing top 5 important NLP-extracted features. The correlation values of the LR, L2-LR, LGBM, and SVM models were more influenced by using the NLP-extracted features. With the replacement of gold standard features, the correlation values gradually increased and exceeded 0.950. The ANN model did not achieve a good concordance correlation even when the top 5 important NLP-extracted features were replaced.

Table 6. Performance of the multiturn question answering model and baseline models.

Feature	BiLSTM[a]-pipeline			BERT[b]-pipeline			BERT-MTQA[c]		
	P[d]	R[e]	F[f]	P	R	F	P	R	F
Tumor density	0.882	0.625	0.732	0.889	0.667	0.762	0.938	0.938	0.938
MLNLA[g]	1.000	0.640	0.780	1.000	0.720	0.837	1.000	0.960	0.980
TLA[h]	0.967	0.892	0.928	0.984	0.938	0.961	0.984	0.954	0.969
Lobulation	0.889	0.533	0.667	0.909	0.667	0.769	1.000	0.867	0.929
TSA[i]	0.967	0.892	0.928	0.984	0.938	0.961	0.984	0.954	0.969
MLNSA[j]	1.000	0.750	0.857	1.000	0.750	0.857	1.000	0.938	0.968
Pleural indentation	0.931	0.818	0.871	0.964	0.818	0.885	1.000	0.848	0.918
Tumor location	0.984	0.897	0.938	0.968	0.897	0.931	0.985	0.985	0.985
Spiculation	1.000	0.727	0.842	1.000	0.773	0.872	1.000	1.000	1.000
Vessel invasion	1.000	0.111	0.200	1.000	0.222	0.364	1.000	0.556	0.714
HLNLA[k]	1.000	0.778	0.875	1.000	0.833	0.909	1.000	1.000	1.000
HLNSA[l]	1.000	0.750	0.857	1.000	0.750	0.857	1.000	1.000	1.000
Average	0.968	0.701	0.790	0.975	0.748	0.830	0.991	0.917	0.948

[a]BiLSTM: bidirectional long short-term memory.

[b]BERT: bidirectional encoder representations from transformers.

[c]MTQA: multiturn question answering.

[d]P: precision.

[e]R: recall.

[f]F: F1 score.

[g]MLNLA: mediastinal lymph node long axis.

[h]TLA: tumor long axis.

[i]TSA: tumor short axis.

[j]MLNSA: mediastinal lymph node short axis.

[k]HLNLA: hilar lymph node long axis.

[l]HLNSA: hilar lymph node short axis.

Table 7. Performance of the multiturn question answering model for feature extraction.

Feature	Accuracy	Precision	Recall	F1 score
Tumor density	0.940	0.875	0.915	0.893
MLNLA[a]	0.965	0.927	0.927	0.927
TLA[b]	0.974	0.974	0.974	0.974
Lobulation	0.923	0.993	0.716	0.832
TSA[c]	0.972	0.972	0.972	0.972
MLNSA[d]	0.986	0.918	0.931	0.924
Pleural indentation	0.917	0.903	0.938	0.920
Tumor location	0.994	0.990	0.990	0.990
Spiculation	0.979	0.988	0.945	0.966
Vessel invasion	0.982	0.932	0.788	0.854
HLNLA[e]	0.965	1.000	0.811	0.896
HLNSA[f]	0.986	0.982	0.848	0.911

[a]MLNLA: mediastinal lymph node long axis.

[b]TLA: tumor long axis.

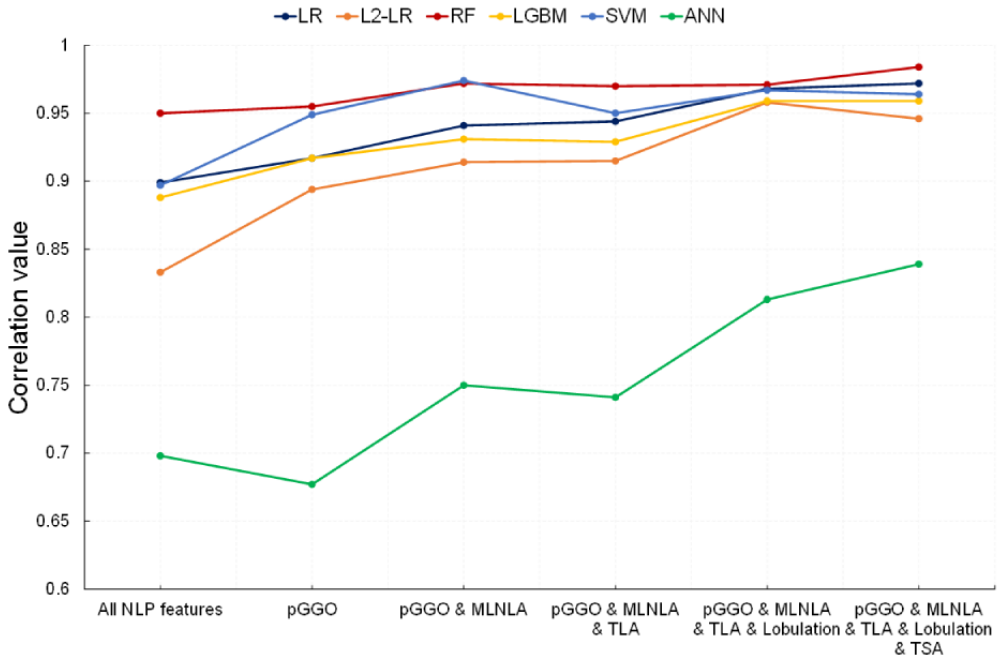
[c]TSA: tumor short axis.

[d]MLNSA: mediastinal lymph node short axis.

[e]HLNLA: hilar lymph node long axis.

[f]HLNSA: hilar lymph node short axis.

Figure 4. Concordance correlation values between pN2 prediction models using complete and partial gold standard features. LR: logistic regression; L2-LR: L2-logistic regression; RF: random forest; LGBM: LightGBM; SVM: support vector machine; ANN: artificial neural network: NLP: natural language processing; pGGO: pure ground glass opacity; MLNLA: mediastinal lymph node long axis; TLA: tumor long axis; TSA: tumor short axis.



Discussion

Principal Findings

In this study, we explored the feasibility of using EMR to develop machine learning models to predict LNM for patients with NSCLC. The important features about the primary tumor and lymph nodes were extracted from the CT reports using NLP technique to support the model development. To the best of our knowledge, this is the first study to use NLP technique to extract features to build preoperative LNM prediction models for patients with NSCLC. Experimental results indicate that the RF model achieved the best performances with 0.792 AUC value and 0.456 AP value for pN2 LNM prediction. All machine

learning models outperformed the size criterion and clinician's evaluation.

Among all models, the LR, L2-LR, RF, and LGBM provide the feature importance to show the connections between the patient features and LNM status. CEA, tumor density, CA125, MLNLA, TLA, lobulation, and TSA were ranked in the top 10 important features by the machine learning models, which was consistent with the results of univariate analysis. Squamous cell carcinoma antigen (SCCAg) was also identified as a top 10 important feature by the models, although univariate analysis did not show significance. However, SCCAg has been proved to be associated with LNM in esophageal squamous cell carcinoma [46], anus squamous cell carcinoma [47], oral-cavity squamous cell carcinoma [48], and cervical squamous cell carcinoma [49]. It is also a poor prognostic factor of lung squamous cell carcinoma and upgrading the patient stage is recommended [50,51]. Surprisingly, TLA was identified as an important feature with negative weight by the LR model, which means the longer the TLA is, the lower the risk of pN2 LNM the patient may have. The result is contrary to the result of univariate analysis and may be caused by multicollinearity or interactions between the features [52]. In the L2-LR model, the TLA was not ranked in the top 10 important features, indicating the L2 regularization can indeed reduce the influence of multicollinearity and improve the interpretability of the model [53]. In addition, other features like right middle lobe cardiovascular disease also suffered interpretability problems, which may be hard to accept in clinical practice. Therefore, more robust interpretable machine learning algorithms are needed to make accurate predictions while giving more reasonable explanations.

In this study, we innovatively extracted features from CT reports and used them to develop LNM prediction models. The concordance correlations between the predicted probabilities of models using NLP-extracted features, partially NLP-extracted features, and gold standard features indicate that the automatically developed models can obtain similar predictive results to those of models using gold standard features. This finding implicitly indicates it is possible to build models using a large amount of unstructured data and update them automatically. More importantly, it can also reduce the burden of manual feature extraction to improve the usability of the prediction models in clinical practice.

Limitations

Although the experimental results show that machine learning models using CT reports, demographic information, medical history, and biomarker data can achieve better performances than the size criterion and clinician's evaluation on the collected data, external validation is still needed to further prove the effectiveness and generalization of the NLP and LNM prediction models. Note that the writing styles of CT reports from different medical centers may vary greatly, which poses a huge challenge to the NLP model developed using the CT reports from a single medical center. Transfer learning is a proper strategy to solve the problem by fine-tuning the model to adapt to CT reports from other centers. Overall, multicenter data is necessary to develop a more robust and generalizable NLP and LNM prediction model.

Furthermore, many studies have proved that there are deep features or radiomics features related to LNM in the CT images [54-60]. Clinicians cannot recognize these with the naked eye, so these features may provide extra information about the metastasis status. In the future, we will extract the image features and combine them with the features in this study to develop more robust, accurate multimodal LNM prediction models.

Conclusions

In this study, we used NLP and machine learning methods to develop the LNM prediction models for patients with NSCLC using EMRs. The RF model achieved the best performance with 0.792 AUC value and 0.456 AP value for pN2 prediction and 0.768 AUC value and 0.524 AP value for pN1&N2 prediction. All machine learning models outperformed the size criterion and clinician's evaluation. Furthermore, the experimental results indicate that the NLP model can effectively extract features from CT reports to support the automatic development and update of the LNM prediction model and may facilitate the application of models in clinical practice.

Acknowledgments

The publication of this paper was funded by grant 2018YFC0910700 from the National Key Research and Development Program of China.

Authors' Contributions

DH, SL, XL, and NW conceptualized the study. SL acquired the clinical data. DH and HZ designed and implemented the algorithms and conducted the experiments. DH, HZ, and SL analyzed the experimental results. DH wrote the manuscript with revision assistance from SL, XL, and NW. All authors have read and approved the manuscript.

Conflicts of Interest

None declared.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 Feb 04:1 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Cancer facts and figures 2021. American Cancer Society. URL: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2021.html [accessed 2021-07-14]
3. Ettinger D, Wood D, Aisner D, Akerley W, Bauman J, Chirieac L, et al. Non-Small Cell Lung Cancer, Version 5.2017, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2017 Apr;15(4):504-535 [FREE Full text] [doi: 10.6004/jnccn.2017.0050] [Medline: 28404761]
4. Hu D, Li S, Huang Z, Wu N, Lu X. Predicting postoperative non-small cell lung cancer prognosis via long short-term relational regularization. Artif Intell Med 2020 Jul;107:101921. [doi: 10.1016/j.artmed.2020.101921] [Medline: 32828458]
5. Detterbeck FC, Boffa DJ, Kim AW, Tanoue LT. The Eighth Edition Lung Cancer Stage Classification. Chest 2017 Jan;151(1):193-203. [doi: 10.1016/j.chest.2016.10.010] [Medline: 27780786]
6. Silvestri GA, Gonzalez AV, Jantz MA, Margolis ML, Gould MK, Tanoue LT, et al. Methods for staging non-small cell lung cancer: diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest 2013 May;143(5 Suppl):e211S-e250S. [doi: 10.1378/chest.12-2355] [Medline: 23649440]
7. Schmidt-Hansen M, Baldwin DR, Zamora J. FDG-PET/CT imaging for mediastinal staging in patients with potentially resectable non-small cell lung cancer. JAMA 2015 Apr 14;313(14):1465-1466. [doi: 10.1001/jama.2015.2365] [Medline: 25871673]
8. Zhang C, Song Q, Zhang L, Wu X. Development of a nomogram for preoperative prediction of lymph node metastasis in non-small cell lung cancer: a SEER-based study. J Thorac Dis 2020 Jul;12(7):3651-3662 [FREE Full text] [doi: 10.21037/jtd-20-601] [Medline: 32802444]
9. Navani N, Fisher DJ, Tierney JF, Stephens RJ, Burdett S, NSCLC Meta-analysis Collaborative Group. The accuracy of clinical staging of stage I-IIIa non-small cell lung cancer: an analysis based on individual participant data. Chest 2019 Mar;155(3):502-509 [FREE Full text] [doi: 10.1016/j.chest.2018.10.020] [Medline: 30391190]
10. Lv X, Wu Z, Cao J, Hu Y, Liu K, Dai X, et al. A nomogram for predicting the risk of lymph node metastasis in T1-2 non-small-cell lung cancer based on PET/CT and clinical characteristics. Transl Lung Cancer Res 2021 Jan;10(1):430-438 [FREE Full text] [doi: 10.21037/tlcr-20-1026] [Medline: 33569324]
11. Chen K, Yang F, Jiang G, Li J, Wang J. Development and validation of a clinical prediction model for N2 lymph node metastasis in non-small cell lung cancer. Ann Thorac Surg 2013 Nov;96(5):1761-1768. [doi: 10.1016/j.athoracsur.2013.06.038] [Medline: 23998401]
12. Miao H, Shaolei L, Nan L, Yumei L, Shanyuan Z, Fangliang L, et al. Occult mediastinal lymph node metastasis in FDG-PET/CT node-negative lung adenocarcinoma patients: risk factors and histopathological study. Thorac Cancer 2019 Jun;10(6):1453-1460 [FREE Full text] [doi: 10.1111/1759-7714.13093] [Medline: 31127706]
13. Verdial FC, Madtes DK, Hwang B, Mulligan MS, Odem-Davis K, Waworuntu R, et al. Prediction model for nodal disease among patients with non-small cell lung cancer. Ann Thorac Surg 2019 Jun;107(6):1600-1606 [FREE Full text] [doi: 10.1016/j.athoracsur.2018.12.041] [Medline: 30710518]
14. Shafazand S, Gould MK. A clinical prediction rule to estimate the probability of mediastinal metastasis in patients with non-small cell lung cancer. J Thorac Oncol 2006 Nov;1(9):953-959 [FREE Full text] [Medline: 17409978]
15. Farjah F, Lou F, Sima C, Rusch VW, Rizk NP. A prediction model for pathologic N2 disease in lung cancer patients with a negative mediastinum by positron emission tomography. J Thorac Oncol 2013 Sep;8(9):1170-1180 [FREE Full text] [doi: 10.1097/JTO.0b013e3182992421] [Medline: 23945387]
16. Song C, Kimura D, Sakai T, Tsushima T, Fukuda I. Novel approach for predicting occult lymph node metastasis in peripheral clinical stage I lung adenocarcinoma. J Thorac Dis 2019 Apr;11(4):1410-1420 [FREE Full text] [doi: 10.21037/jtd.2019.03.57] [Medline: 31179083]
17. Yim W, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol 2016 Jun 01;2(6):797-804. [doi: 10.1001/jamaoncol.2016.0213] [Medline: 27124593]
18. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
19. Monteiro M, Fonseca AC, Freitas AT, Pinho E Melo T, Francisco AP, Ferro JM, et al. Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM Trans Comput Biol Bioinform 2018;15(6):1953-1959. [doi: 10.1109/TCBB.2018.2811471] [Medline: 29994736]
20. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open 2020 Jan 03;3(1):e1918962 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.18962] [Medline: 31922560]
21. Ali F, El-Sappagh S, Islam S, Kwak D, Ali A, Imran M. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf Fusion 2020;63:208-222 [FREE Full text] [doi: 10.1016/j.inffus.2020.06.008]

22. Datta S, Bernstam EV, Roberts K. A frame semantic overview of NLP-based information extraction for cancer-related EHR notes. J Biomed Inform 2019 Dec;100:103301 [FREE Full text] [doi: 10.1016/j.jbi.2019.103301] [Medline: 31589927]
23. Si Y, Roberts K. A frame-based NLP system for cancer-related information extraction. AMIA Annu Symp Proc 2018;2018:1524-1533 [FREE Full text] [Medline: 30815198]
24. Yim W, Denman T, Kwan SW, Yetisgen M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. AMIA Jt Summits Transl Sci Proc 2016;2016:455-464 [FREE Full text] [Medline: 27570686]
25. Savova GK, Tseytlin E, Finan S, Castine M, Miller T, Medvedeva O, et al. DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. Cancer Res 2017 Nov 01;77(21):e115-e118 [FREE Full text] [doi: 10.1158/0008-5472.CAN-17-0615] [Medline: 29092954]
26. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med 2016 Jan;66:29-39 [FREE Full text] [doi: 10.1016/j.artmed.2015.09.007] [Medline: 26481140]
27. Hu D, Zhang H, Li S, Wang Y, Wu N, Lu X. Automatic extraction of lung cancer staging information from computed tomography reports: deep learning approach. JMIR Med Inform 2021 Jul 21;9(7):e27955 [FREE Full text] [doi: 10.2196/27955] [Medline: 34287213]
28. Zheng C, Huang BZ, Agazaryan AA, Creekmur B, Osuj TA, Gould MK. Natural language processing to identify pulmonary nodules and extract nodule characteristics from radiology reports. Chest 2021 Nov;160(5):1902-1914. [doi: 10.1016/j.chest.2021.05.048] [Medline: 34089738]
29. Sugimoto K, Takeda T, Oh J, Wada S, Konishi S, Yamahata A, et al. Extracting clinical terms from radiology reports with deep learning. J Biomed Inform 2021 Apr;116:103729 [FREE Full text] [doi: 10.1016/j.jbi.2021.103729] [Medline: 33711545]
30. Chen L, Song L, Shao Y, Li D, Ding K. Using natural language processing to extract clinically useful information from Chinese electronic medical records. Int J Med Inform 2019 Apr;124:6-12. [doi: 10.1016/j.ijmedinf.2019.01.004] [Medline: 30784428]
31. Martinez D, Pitson G, MacKinlay A, Cavedon L. Cross-hospital portability of information extraction of cancer staging information. Artif Intell Med 2014 Sep;62(1):11-21. [doi: 10.1016/j.artmed.2014.06.002] [Medline: 25001545]
32. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 2017 Dec;69:177-187 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.011] [Medline: 28428140]
33. Bozkurt S, Lipson JA, Senol U, Rubin DL. Automatic abstraction of imaging observations with their characteristics from mammography reports. J Am Med Inform Assoc 2015 Apr;22(e1):e81-e92. [doi: 10.1136/amiajnl-2014-003009] [Medline: 25352567]
34. Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. J Biomed Inform 2016 Aug;62:224-231 [FREE Full text] [doi: 10.1016/j.jbi.2016.07.001] [Medline: 27388877]
35. Sui X, Liu T, Huang Q, Hou Y, Wang Y, Kang G, et al. P2.09-29 Automatic lung cancer staging from medical reports using natural language processing. J Thor Oncol 2018 Oct;13(10):S772. [doi: 10.1016/j.jtho.2018.08.1326]
36. Yuan Q, Cai T, Hong C, Du M, Johnson BE, Lanuti M, et al. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA Netw Open 2021 Jul 01;4(7):e2114723 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.14723] [Medline: 34232304]
37. Li X, Yin F, Sun Z, Li X, Yuan A, Chai D, et al. Entity-relation extraction as multi-turn question answering. 2019 Presented at: Proc 57th Annu Meet Assoc Comput Linguist; 2019; Florence p. 1340-1350. [doi: 10.18653/v1/p19-1129]
38. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Arxiv. Preprint posted online Oct 10, 2018 2018:1 [FREE Full text]
39. Hosmer D, Lemeshow S, Sturdivant R. Applied Logistic Regression. 3rd ed. Hoboken: John Wiley & Sons; 2013.
40. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970 Feb;12(1):55-67. [doi: 10.1080/00401706.1970.10488634]
41. Breiman L. Random forests. Mach Learn 2001;45(1):5-32 [FREE Full text] [doi: 10.1023/A:1010933404324]
42. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 2017 Presented at: 31st Conf Neural Inf Process Syst (NIPS 2017); 2017; Long Beach URL: https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
43. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
44. Jain A, Mao J, Mohiuddin K. Artificial neural networks: a tutorial. Computer (Long Beach Calif) 1996;29(3):31-44. [doi: 10.1109/2.485891]
45. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans Audio Speech Lang Process 2021;29:3504-3514. [doi: 10.1109/taslp.2021.3124365]
46. Shimada H, Nabeya Y, Okazumi S, Matsubara H, Shiratori T, Gunji Y, et al. Prediction of survival with squamous cell carcinoma antigen in patients with resectable esophageal squamous cell carcinoma. Surgery 2003 May;133(5):486-494. [doi: 10.1067/msy.2003.139] [Medline: 12773976]

47. Williams M, Swampillai A, Osborne M, Mawdsley S, Hughes R, Harrison M, Mount Vernon Colorectal Cancer Network. Squamous cell carcinoma antigen: a potentially useful prognostic marker in squamous cell carcinoma of the anal canal and margin. Cancer 2013 Jul 01;119(13):2391-2398 [FREE Full text] [doi: 10.1002/cncr.28055] [Medline: 23576077]
48. Lin W, Chen I, Wei F, Huang J, Kang C, Hsieh L, et al. Clinical significance of preoperative squamous cell carcinoma antigen in oral-cavity squamous cell carcinoma. Laryngoscope 2011 May;121(5):971-977. [doi: 10.1002/lary.21721] [Medline: 21520110]
49. Xu D, Wang D, Wang S, Tian Y, Long Z, Ren X. Correlation between squamous cell carcinoma antigen level and the clinicopathological features of early-stage cervical squamous cell carcinoma and the predictive value of squamous cell carcinoma antigen combined with computed tomography scan for lymph node metastasis. Int J Gynecol Cancer 2017 Nov;27(9):1935-1942. [doi: 10.1097/IGC.0000000000001112] [Medline: 28914639]
50. Kinoshita T, Ohtsuka T, Yotsukura M, Asakura K, Goto T, Kamiyama I, et al. Prognostic impact of preoperative tumor marker levels and lymphovascular invasion in pathological stage I adenocarcinoma and squamous cell carcinoma of the lung. J Thorac Oncol 2015 Apr;10(4):619-628 [FREE Full text] [doi: 10.1097/JTO.0000000000000480] [Medline: 25634009]
51. Kinoshita T, Ohtsuka T, Hato T, Goto T, Kamiyama I, Tajima A, et al. Prognostic factors based on clinicopathological data among the patients with resected peripheral squamous cell carcinomas of the lung. J Thorac Oncol 2014 Dec;9(12):1779-1787 [FREE Full text] [doi: 10.1097/JTO.0000000000000338] [Medline: 25226427]
52. Tolles J, Meurer WJ. Logistic regression: relating patient characteristics to outcomes. JAMA 2016 Aug 02;316(5):533-534. [doi: 10.1001/jama.2016.7653] [Medline: 27483067]
53. Marquardt DW, Snee RD. Ridge regression in practice. Am Statistician 1975 Feb;29(1):3-20. [doi: 10.1080/00031305.1975.10479105]
54. Gu Y, She Y, Xie D, Dai C, Ren Y, Fan Z, et al. A texture analysis-based prediction model for lymph node metastasis in stage Ia lung adenocarcinoma. Ann Thorac Surg 2018 Jul;106(1):214-220. [doi: 10.1016/j.athoracsur.2018.02.026] [Medline: 29550204]
55. Hosny A, Parmar C, Quackenbush J, Schwartz LH. Artificial intelligence in radiology. Nat Rev Cancer 2018 Dec;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
56. Cong M, Feng H, Ren J, Xu Q, Cong L, Hou Z, et al. Development of a predictive radiomics model for lymph node metastases in pre-surgical CT-based stage IA non-small cell lung cancer. Lung Cancer 2020 Jan;139:73-79 [FREE Full text] [doi: 10.1016/j.lungcan.2019.11.003] [Medline: 31743889]
57. Zhao X, Wang X, Xia W, Li Q, Zhou L, Li Q, et al. A cross-modal 3D deep learning for accurate lymph node metastasis prediction in clinical stage T1 lung adenocarcinoma. Lung Cancer 2020 Jul;145:10-17. [doi: 10.1016/j.lungcan.2020.04.014] [Medline: 32387813]
58. Wang X, Nan W, Yan S, Li Q, Guo N, Guo Z. MA05.11 radiomics analysis using SVM predicts mediastinal lymph nodes status of squamous cell lung cancer by pre-treatment chest CT scan. J Thor Oncol 2018 Oct;13(10):S374. [doi: 10.1016/j.jtho.2018.08.357]
59. He L, Huang Y, Yan L, Zheng J, Liang C, Liu Z. Radiomics-based predictive risk score: a scoring system for preoperatively predicting risk of lymph node metastasis in patients with resectable non-small cell lung cancer. Chin J Cancer Res 2019 Aug;31(4):641-652 [FREE Full text] [doi: 10.21147/j.issn.1000-9604.2019.04.08] [Medline: 31564807]
60. Yoo J, Cheon M, Park YJ, Hyun SH, Zo JI, Um S, et al. Machine learning-based diagnostic method of pre-therapeutic F-FDG PET/CT for evaluating mediastinal lymph nodes in non-small cell lung cancer. Eur Radiol 2021 Jun;31(6):4184-4194. [doi: 10.1007/s00330-020-07523-z] [Medline: 33241521]

Abbreviations

ANN: artificial neural network
AP: average precision
AUC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
BiLSTM: bidirectional long short-term memory
BI-RADS: breast imaging-reporting and data system
CA125: carbohydrate antigen 12-5
CEA: carcinoembryonic antigen
cN: clinical N stage
EMR: electronic medical record
LGBM: LightGBM
LNM: lymph node metastasis
LR: logistic regression
L2-LR: L2-logistic regression
MLNLA: mediastinal lymph node long axis
MTQA: multiturn question answering

NLP: natural language processing
NSCLC: non–small cell lung cancer
NSE: neuron-specific enolase
PET-CT: positron emission tomography–computed tomography
pN: pathological N stage
PR: precision-recall curve
RF: random forest
ROC: receiver operating characteristic curve
SCCAg: squamous cell carcinoma antigen
SUVmax: maximum standardized uptake value
SVM: support vector machine
TLA: tumor long axis
TSA: tumor short axis

Edited by C Lovis; submitted 22.12.21; peer-reviewed by YH Kim, V Rajan; comments to author 27.03.22; revised version received 31.03.22; accepted 11.04.22; published 25.04.22.

Please cite as:
Hu D, Li S, Zhang H, Wu N, Lu X
Using Natural Language Processing and Machine Learning to Preoperatively Predict Lymph Node Metastasis for Non–Small Cell Lung Cancer With Electronic Medical Records: Development and Validation Study
JMIR Med Inform 2022;10(4):e35475
URL: https://medinform.jmir.org/2022/4/e35475
doi:10.2196/35475
PMID:35468085

Original Paper

Investigating Health Context Using a Spatial Data Analytical Tool: Development of a Geospatial Big Data Ecosystem

Timothy Haithcoat[1], MS; Danlu Liu[1], MSc; Tiffany Young[1], MSc; Chi-Ren Shyu[1], PhD

Institute for Data Science and Informatics, University of Missouri, Columbia, MO, United States

Corresponding Author:
Chi-Ren Shyu, PhD
Institute for Data Science and Informatics
University of Missouri
22 Heinkel Building
Columbia, MO, 65211
United States
Phone: 1 573 882 3884
Fax: 1 573 884 8709
Email: shyuc@missouri.edu

Abstract

Background: Enabling the use of spatial context is vital to understanding today's digital health problems. Any given location is associated with many different contexts. The strategic transformation of population health, epidemiology, and eHealth studies requires vast amounts of integrated digital data. Needed is a novel analytical framework designed to leverage location to create new contextual knowledge. The Geospatial Analytical Research Knowledgebase (GeoARK), a web-based research resource has robust, locationally integrated, social, environmental, and infrastructural information to address today's complex questions, investigate context, and spatially enable health investigations. GeoARK is different from other Geographic Information System (GIS) resources in that it has taken the layered world of the GIS and flattened it into a big data table that ties all the data and information together using location and developing its context.

Objective: It is paramount to build a robust spatial data analytics framework that integrates social, environmental, and infrastructural knowledge to empower health researchers' use of geospatial context to timely answer population health issues. The goal is twofold in that it embodies an innovative technological approach and serves to ease the educational burden for health researchers to think spatially about their problems.

Methods: A unique analytical tool using location as the key was developed. It allows integration across source, geography, and time to create a geospatial big table with over 162 million individual locations (X-Y points that serve as rows) and 5549 attributes (represented as columns). The concept of context (adjacency, proximity, distance, etc) is quantified through geoanalytics and captured as new distance, density, or neighbor attributes within the system. Development of geospatial analytics permits contextual extraction and investigator-initiated eHealth and mobile health (mHealth) analysis across multiple attributes.

Results: We built a unique geospatial big data ecosystem called GeoARK. Analytics on this big table occur across resolution groups, sources, and geographies for extraction and analysis of information to gain new insights. Case studies, including telehealth assessment in North Carolina, national income inequality and health outcome disparity, and a Missouri COVID-19 risk assessment, demonstrate the capability to support robust and efficient geospatial understanding of a wide spectrum of population health questions.

Conclusions: This research identified, compiled, transformed, standardized, and integrated multifaceted data required to better understand the context of health events within a large location-enabled database. The GeoARK system empowers health professionals to engage more complex research where the synergisms of health and geospatial information will be robustly studied beyond what could be accomplished today. No longer is the need to know how to perform geospatial processing an impediment to the health researcher, but rather the development of how to think spatially becomes the greater challenge.

(JMIR Med Inform 2022;10(4):e35073) doi:10.2196/35073

KEYWORDS

context; Geographic Information System; big data; equity; population health; public health; digital health; eHealth; location; geospatial; data analytics; analytical framework; medical informatics; research knowledgebase

Introduction

Health researchers need integrated social, environmental, and infrastructural information to extend the scope of health care and address the complex questions and contextual relationships surrounding health outcomes. Any given location is associated with different contexts—physical, biological, environmental, infrastructural, economic, social, and cultural—all of which can affect population health, disease risk, and access to health care. Geographic context plays a growing role in connecting heterogeneous geoenabled information, especially in health research [1-4]. Spatial context includes elements and interactions with both the societal and the physical infrastructures associated with an individual's daily activities. This includes accessibility, surrounding natural and built environments, social behaviors, and any related location-specific exposures, understanding that these elements change across geographic areas, scales, and time. Impactful health research that can be applied to real-world issues and problems must be grounded within the context of place [5-7]. Location and the location's context both matter [8-10]!

The strategic transformation of population health, epidemiology, and eHealth studies require vast amounts of integrated digital data to create understanding that can then support decisions [11]. Questions asked today are more complex than ever before, implicitly tied to understanding context [12-18]. Health researchers have used the Geographic Information System (GIS) to identify, mitigate, and address a myriad of factors affecting health disparities [19-22], health assessments [23-26], health-environment interactions [27-32], health-cultural interactions [33-35], and health service access [36-42]. GIS analysis is expanding within health analysis, but its use is often focused on thematic single-variable maps and their visualization [43-45]. Medical researchers who study health disparities tend to focus on demographic, social, or economic variables from local to national levels, both cross-sectional and over time, that are available from the decennial census or the American Community Survey (ACS). Although there are exceptions [46], far fewer use variables related to the natural, physical, or built environment, primarily because they are more challenging to obtain.

Although advancement is evident in the various web-mapping sites across the federal health realm (the Centers for Disease Control [CDC] and Prevention's Heart Disease and Stroke Maps, the National Institutes of Health [NIH] and National Cancer Institute's Cancer Atlas and state profiles, and the Environmental Protection Agency's [EPA] EnviroAtlas), several issues persist. Although integrated information sources available for researchers are growing [47-49], they each portray only a specific view of that entity's mandated purview. Most provide visualization of singular attributes at a time and rely on the user to mentally synthesize these pieces of information to generate understanding. It remains a challenge for health researchers to locate and evaluate what specific attributes exist and at what geographies. Moreover, many health researchers "don't know what they don't know" with regard to geospatial data. The ability to create new hypotheses is missed if researchers are not aware of the availability of data or the types of questions that could be posed that could further expand their research. More importantly, development of contextual relationships among variables could be discovered through spatial analytics. In addition, the quantification of interactions of health, demographics, infrastructure, and environmental elements in terms of interplay and synergy remains elusive.

Needed is a novel analytical framework designed to leverage location to create new contextual knowledge and associations among otherwise disjointed data. This would aid evidence-based exploration of relationships among layers, discover patterns of interaction, and support clinical sampling designs where quantification and location are interwoven. This paper outlines a new big data approach to building and evolving such a geoenabled health information system. The Geospatial Analytical Research Knowledgebase (GeoARK) is an informatics and data science solution that uses advanced complex contextual queries across multiresolution locational information to geoenable health research.

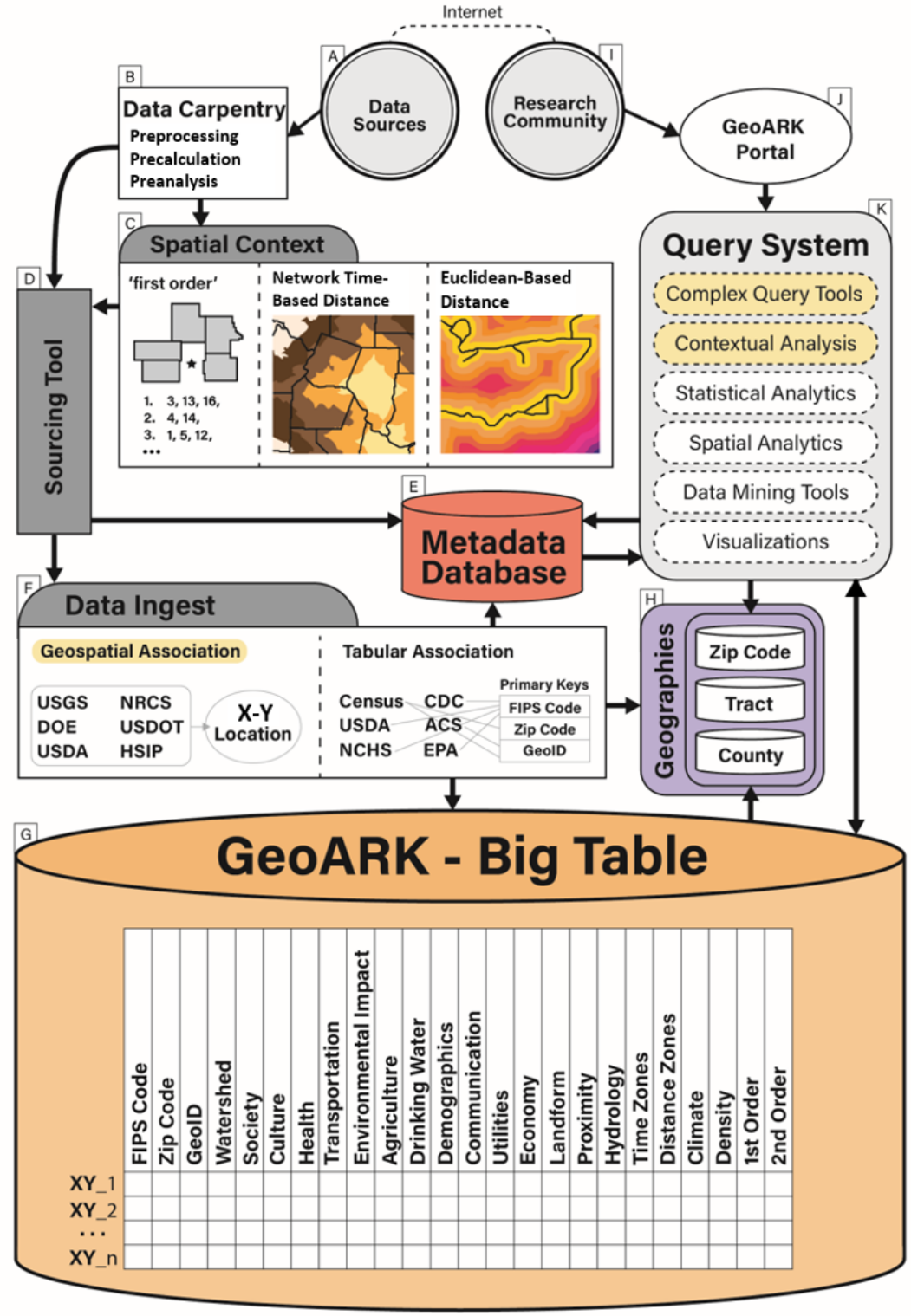
The objective of GeoARK is to transform attitudes and empower health research where real-world problems are examined in geoenabled context. We can gain efficiencies through integrated heterogeneous public information sources and the establishment of context through geospatial measures, such as proximity, adjacency, network analysis, and spatial analysis. These then form a new complex of attributes within a single geoenabled knowledgebase. It can support a broad spectrum of health research, including health disparities, telemedicine, communicable disease management, zoonotic disease surveillance, environmental health, and health access policy making. It enables eHealth researchers to bring their own collection of eHealth or mobile health (mHealth) events and have user-selected attribute data compiled at those points or output artificial intelligence/machine learning (AI/ML)-friendly databases for further analysis. The contextualization of existing research would enhance the scope of that research.

Methods

GeoARK Design

This paper describes GeoARK and its potential to greatly extend eHealth research. It outlines how the system was designed and demonstrates how its design leads to its potential within health research. The GeoARK system (Figure 1) comprises multiple components that interact to form a complete process for the integration, documentation, and spatial registration of data into a single queryable big table that we call GeoARK-Big Table (GeoARK-BT) in this paper. It can be used by health researchers to accelerate the use of spatial data and exploit local context within analyses.

Figure 1. Geospatial Analytical Research Knowledgebase (GeoARK) System design.



The actual spatial framework of GeoARK-BT is based on a dense distribution of points across the United States. The spatial base is a hexagon tessellation of points blanketing the United States at a spacing of 161 m (1/10th of a mile, or 528 ft). Centroids of census blocks with an area less than 67,261 m^2 (16.6 acres) are integrated into the tessellation to better capture features in more densely populated areas. Proximal polygons are calculated for each point that allows for area totals as well as aggregation into user-specified geographies to occur. These units form a coherent framework for cataloging data over geographical space. The point locations create the sampling framework through which GeoARK captures and encodes the locational variability that exists across the databases integrated. For the United States, there are 162 million points with basic information (5549 attributes) in our current system. Each point is a row, with all attributes associated with that location becoming columns in the database, while each attribute is a column with 162 million rows, with each element of the column

representing a specific location's attribute. The points are stored in a Hadoop Distributed File System (HDFS). The total size is 12.5 TB. The data loading process for all 5549 attributes across the 50 states took 2585 min using a Dell PowerEdge R740xdcompute node with a dual Intel(R) Xeon(R) Gold 6138 CPU (80 cores) and 384 GB memory.

Data Sourcing and Metadata

The GeoARK system integrates interdisciplinary public data existing in a wide variety of formats (tabular, raster, point, line, and polygon). This increases the efficiency of research since many data elements and sources are challenging for researchers to compile and uniformly integrate for analysis. The GeoARK-BT includes demographic, social, economic, educational, cultural, infrastructure, and environmental attributes from a growing variety of sources, as listed in Multimedia Appendix 1.

A sourcing tool was developed to standardize the collection, documentation, and logging of each data source being added into the GeoARK-BT collection to ensure data quality and integrity for long-term tracking and maintenance. Once a data source is identified, it is added to the GeoARK source table and then data set information is collected and compiled into the data set descriptive listing (ie, data use agreements, constraints, URLs). The metadata database then catalogs and records individual attribute information from these data sets. The metadata database includes sources, metadata (for both data sets and their associated attributes), and attribute links for the GeoARK-BT. The NIH's Findable, Accessible, Interoperable, Reusable (FAIR) initiative [50] provides a use area for this metadata. Data added to the big table use the attribute lookup table to set attribute field names. Data sources and attribute fields have also been assigned to an International Organization for Standardization (ISO) 19115 thematic category [51]. Natural language tags describing each attribute were also added. Once attributes are loaded, these metadata elements facilitate discovery, query, crediting, and reuse, with all metadata fields being searchable using MongoDB Query Language. Once attribute selection is performed by the researcher, and a data extract is created, a report summarizing the data source information for all data elements contained in the selection is generated. This facilitates the methodological aspects of data collection and documentation for researchers.

Data Ingestion

Relevant open data sources are ingested to the GeoARK system as tabular information or as relative geographic locations. Although these data independently have great singular value, combining these data, using location as the linkage between data sets, is the power of geospatial analysis and the underpinning for the GeoARK system. Data carpentry and preprocessing are required for some data sources and elements. Attributes being used as links need to be standardized, and categorical data need to be transformed. In some cases, new derived attributes are calculated through aggregation of existing attributes. Precalculations of percentages, densities, means, quantile breaks, and the results of spatial-based analyses further extend the database. By transforming the raw numeric counts into density measures (ie, population, race, ethnicity, or other density per km^2), we can then tally the points and their areas that are within the area of interest or meet a selection criterion, and derive estimated values for these attributes. This can be accomplished without the need for standard spatial layer intersection procedures where calculation of crossing vectors is required. The process is simply a point in polygon selection. This process allows GeoARK great flexibility in context quantification for applied digital health research.

Tabular data linkage was obtained by a common attribute. Each GeoARK-BT point has been identified as being within a specific Census 2010 block, Census 2020 block, 5-digit zip code, and specific watershed code. Any data sharing a common key could then be added. Data collected at a native geographic level, such as county, zip code, or tract, are loaded directly. For information cataloged at finer units such as block groups and blocks, the associated data are loaded into the GeoARK-BT using the appropriate census link for each point. Scripts for ingestion to, or update of, the GeoARK-BT for recurring data sources (ie, ACS updates) include extract, transform, and load processes for these sources. The data are synced with the GeoARK system to add new, updated, or changed elements.

For geospatial data, linkage was obtained by the X-Y location. Line-based spatial data, such as road networks, and point-based data, such as hospitals, nursing homes, and public health clinics, have been integrated within the GeoARK-BT. To do so, these files needed processing so as to align with the GeoARK points. Data that were spatially analyzed for contextual measures (buffers, Euclidean distance, network time, etc) were converted into polygon form or a raster representation. These layers were then associated with each GeoARK-BT point and the travel time or distance for the feature assigned. Some data may be categorical (ie, land cover or soils) or continuous in nature (ie, elevation or precipitation), further effecting ease of integration. Such files were directly assessed against the GeoARK-BT proximal polygon representation to generate a series of attributes that capture the values' variability at that location for these data types.

Context Measures

An innovative aspect of GeoARK is that it has precalculated spatial context measures for many features. The simplest contextual measure is presence within a geography or gridded cell. In another form, context is represented as proximity between a location and features of interest (ie, distance from the stroke unit). It can also take the form of a distance from a linear object (ie, power lines). Proximity can also be derived from network modeling to obtain measures of remoteness, isolation, and accessibility (ie, time or distance).

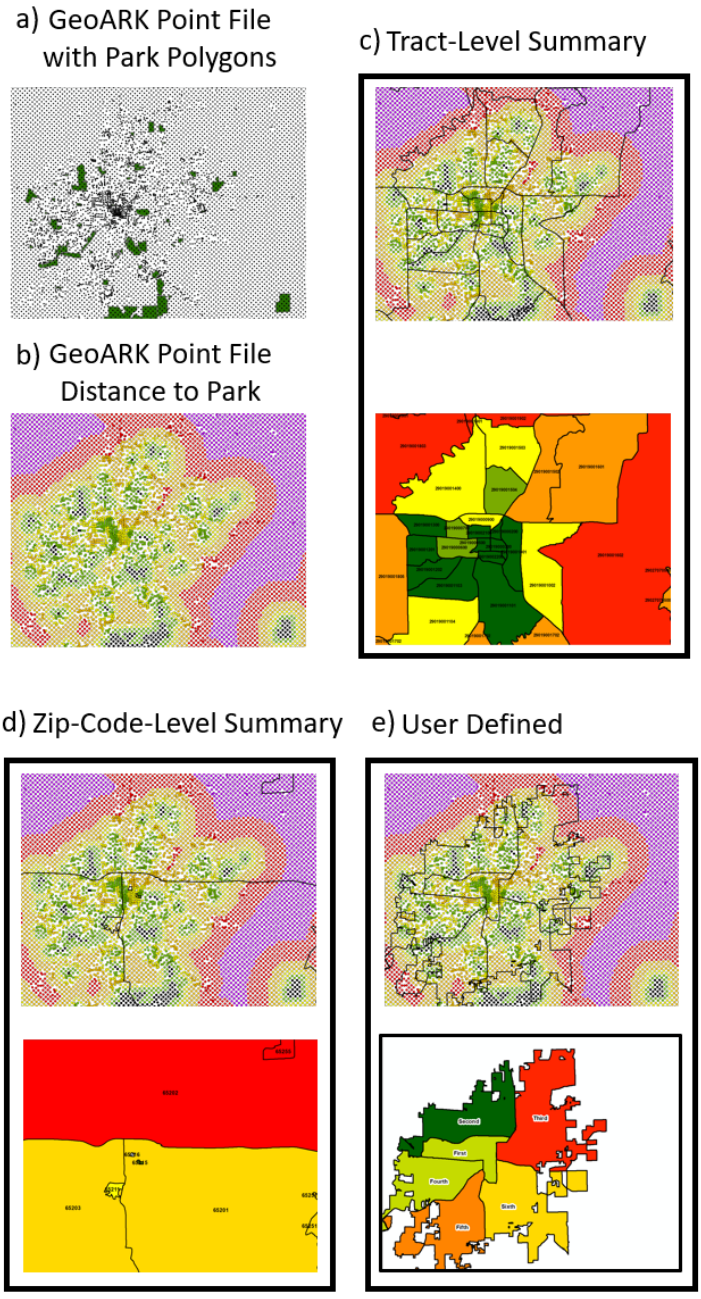
Density measures utilize a grid or distance to tally the number of points, total length of lines, etc, to generate per area metrics. Data such as block-level population, transmission lines, railroads, confined animal feeding operations, and drinking water wells would be cataloged into artificial grids for this density mapping.

Context is also quantified by identifying first- and second-order spatial relationships within geographic levels. These can be thought of as adjacent neighbors and are identified using spatial

analytics. For a given county, the first order is all counties adjacent to this base county. The second order for that same county is all the counties that are adjacent to the first-order counties. These attributes quantify geospatial adjacencies that health researchers can exploit.

Finally, geographic summary levels that are commonly used in research and mapping are precompiled. Although many attributes can be directly related using relational joins at a specific geographic level (county, tract, zip code), other attributes such as distance measures, land cover, elevation, and climate need to be aggregated from the GeoARK-BT points to generate a summary attribute (ie, mean distance to parks) from these features for any region (Figure 2).

Figure 2. This example shows how GeoARK point processing based on a single attribute (Distance to Parks) can be used to generate summaries at various geographic levels. a) Shows GeoARK point layer with parks data superimposed. b) Shows GeoARK points colorized to show distance from parks inherent in their attribution. c-e) Show dark outlines of tract, zip code, and user defined interest - voting wards (respectively) superimposed on the colorized GeoARK points and below each is their resulting geographic summary for mean Distance to Parks. GeoARK: Geospatial Analytical Research Knowledgebase.



GeoARK Utility

The collection, integration, and use of diverse data are foundational to answer today's health problems. Significant disparities exist and can vary across scales from blocks to neighborhoods to regions [3]. In addition, a complex myriad of factors that can affect disparities also exists [52]. In rural contexts [19,36], aging populations, health care access [12,13], sparse populations, environmental exposures [14,15,28], and infrastructure [20] are proven critical factors. In urban contexts, food-deserts [21], crime density and stress [53], and pollution (air, water, light, and noise) [16,54] play possible roles. How

do these factors interact? At what scale are these associations important? Where are these findings located, and are they clustered?

The proposed web portal will provide tools to enable and catalyze a health researcher's ability to move their question into the spatial realm and analyze their area of interest against the broad spectrum of data within the GeoARK-BT. An investigator's area of interest could be an actual physical area (ie, neighborhood, zip code, place) or a collection of health events as X-Y coordinate pairs with which to associate GeoARK attributes. Complex queries can be used to create and refine data extracts that focus on a researcher's question of interest.

To support research, flexible access and powerful interrogation of the GeoARK-BT are required. One of the major strengths of GeoARK is the streamlining of access across data sources and the provision of complex analytic query across multiple timestamps and sources. The query is simply a projection on selected columns within the GeoARK-BT using MongoDB. Indexes were built off-line on each attribute to allow for more efficient on-demand retrieval of information. A single-attribute index takes 623 min to build, and a composite index with 5 attributes takes 791 min. Each index, respectively, has, on average, a 0.66 and 1.05 GB memory footprint for a single and a composite index. Open source analytical tools are to be added to provide further analytical functionality to include descriptive, exploratory, inferential, causal, and predictive approaches to targeted spatial analytical research as GeoARK matures. Points can be selected based on user-defined areas of interest and then aggregated to create a surrogate representation of that area and used to extract user-selected attributes from GeoARK to create a subset for further analysis. The design leverages a big data table where we can have high throughput for data transactions.

The 7 query types, listed in Table 1, range from simple attribute selection to queries that utilize the distance to or from a specific feature type to those that require network travel time or distance. Others might include multitemporal queries concerning what has changed since a particular event or point in time. Still others inquire about features and elements around a particular place or location and the associations found between those factors. Finally, other queries can be built to determine or assess how scale or geographic extent may impact conclusions. Output from each of these types of queries can produce AI/ML-ready data sets leveraging GeoARK's spatial bins and analytical associations.

There is no equivalent system currently available with which to provide side-by-side analytics. When the times presented are compared to the time savings a researcher would obtain through the system's integrated and spatially contextualized information, they provide great value. In addition, through further testing of indexing schemas and optimization of query and search designs, these times are expected to decrease.

Table 1. Examples of query types and their run times when executed against the national GeoARK[a] database. These can range from national to local studies. The first 3 query examples are standard selections based on attribute values or thresholds. The next 3 query examples illustrate the use of the unique spatial dimensional attributes added through the GeoARK system to provide greater geoanalytical power to selections. The final example demonstrates GeoARK's ability to select contextual elements that surround another feature of interest.

Query type	Query example	Query time (min)
Simple geography	Select all records for the state of Missouri, Federal Information Processing Standard (FIPS) code=29.	13.03
Simple variable	Select all county records with a nonmetro flag (2013)=1 in Missouri.	5.81
Complex variable	Black/African American % of total population of zip code >30% AND % total population in poverty >15% AND % households with a single female head of household with children under 18 years of age receiving food stamps >5%.	5.47
Density	Select points with a road density greater than 1500 m (4921 ft) per square kilometer.	0.17
Proximity	Select points with a distance to closest park greater than 400 m (1312 ft).	0.14
Travel time	Select points having 15 min or less travel time to the nearest hospital.	0.07
Contextual	Given a cluster of 3 counties with high cancer incidence, compile and extract all surrounding counties' exposome variables associated with those locations.	1.23

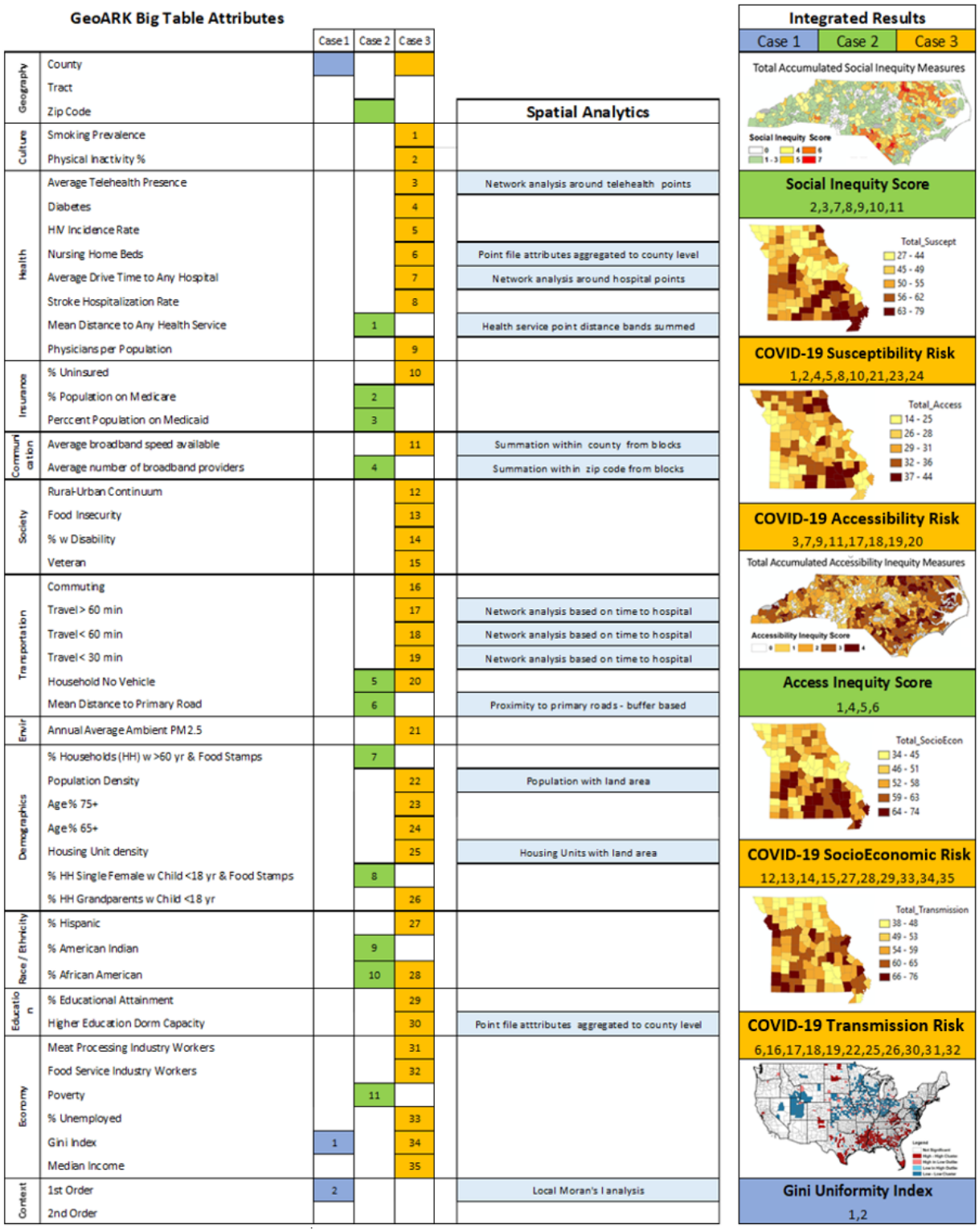
[a]GeoARK: Geospatial Analytical Research Knowledgebase.

Results

Case Studies

Results are presented as 3 case studies (Figure 3) utilizing the GeoARK system. The case studies cover (1) the development of new uniformity measures providing insight into health outcomes (blue), (2) telehealth program evaluation of both growth and impact on rural health access and equity (green), and (3) the development of COVID-19 risk factor assessments (orange). These examples demonstrate the GeoARK system's utility and practical application in support of health research questions.

Figure 3. Case study examples of GeoARK attributes, processes, and outcomes. Each case study is given a color. Attributes used in their respective analyses are likewise color coded. Those attributes derived from spatial analytics are further noted. GeoARK: Geospatial Analytical Research Knowledgebase.



Case Study 1: Associations - Gini Index Example - Geography: County and State

Complex questions: How do health outcomes relate to geographic clustering of income inequality within the United States? Can a new spatial uniformity measure be created to aid understanding of income inequality?

This case study merged geospatial analytics with big data approaches to maximizing the use of the Behavioral Risk Factor Surveillance System (BRFSS) [55] and the ACS [56]. This study examined the income inequality hypothesis using cluster and outlier spatial analysis [57]. We applied this geospatial approach to create 3 innovative measures that captured uniformity in income inequality. We examined the ways that

the Gini coefficient and 3 new spatial uniformity measures were associated with health outcomes. Specifically, the uniformity measures capture the extent to which (1) inequality is uniformly distributed spatially in states regardless of whether the level is high or low, (2) the extent to which states are more uniformly high in inequality across space, and (3) the extent to which they are more uniformly low in inequality. We conclude that residents of states that have more uniformly high inequality across space are more likely to report worse outcomes across several health measures. This case study showed that geospatial big data approaches can extend research on public health topics involving traditional survey data [58]. This also demonstrated how even 1 variable (in this case the Gini index), when spatially analyzed, can create new and useful insights into health investigations and their interpretation.

Case Study 2: Health Equity - Telemedicine Program Reach - Geography (Zip Code Tabulation Areas)

Complex questions: Who and where are the most vulnerable populations in terms of social inequity? Does the telehealth program address these vulnerable populations?

Utilizing aggregated telehealth use data, this case study evaluated a telehealth program's reach, growth, and potential to address equity issues in rural areas. Significant inequities exist and can vary across scales from blocks to neighborhoods to regions [18]. From the occurrence data, the demand for receiving care via the program steadily increased over the 4 quarters, especially in rural areas. Three geospatially based health measures were created to assess and describe context: the social inequity score, the access inequity score, and a combined inequity score. In total, 11 measures, including social determinants (n=7, 64%) and access measures (n=4, 36%), were compiled from 5 sources and tabulated at the zip code level. GeoARK permits selection of both social elements and infrastructure-related accessibility elements. The social elements were pulled from multiple census sources, while the accessibility measures were created through geoanalytics and compiled into zip code boundaries for comparison. To assess the overall context of the delineated reach of the program, a mean combined inequity score was calculated for each zip code and for all zip codes. In zip codes where telemedicine encounters occurred, the population served had higher levels of social inequity and lower access in comparison to both state and rural levels. This telehealth program assessment of health inequity and access in rural regions demonstrated the program's promising reach to vulnerable populations, as associated with the social and accessibility factors measured. These results supported maintaining and continued development of policies for affordable and on-demand telemedicine programs for providing care to rural populations facing inequities [26].

Case Study 3: Population Health - COVID-19 Risk - Geography (County)

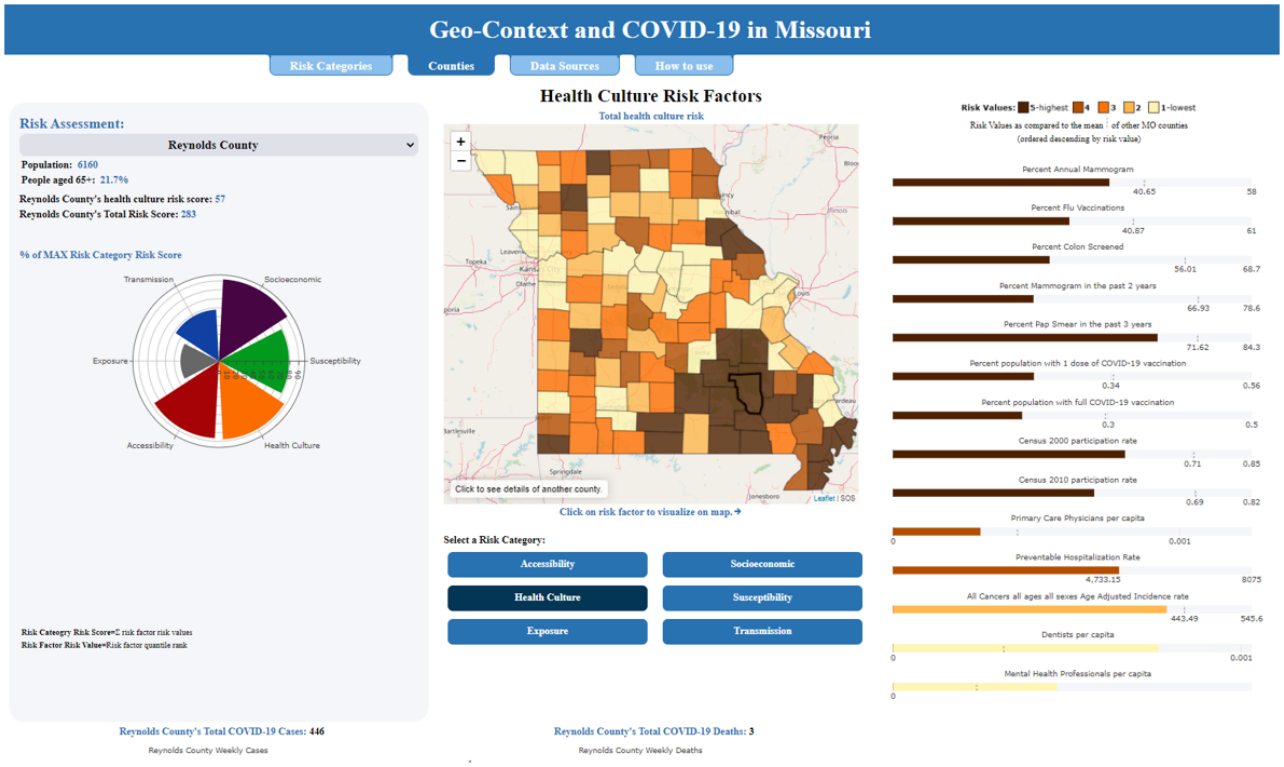
Complex questions: What are the magnitudes of select risk factors, and where are they most prevalent in Missouri? What are the areas of compounded impacts, and do they cluster?

Studies of this COVID-19 pandemic require vast amounts of integrated data to create understanding that can then support decisions. We utilized GeoARK to extract and create 6 distinct thematic risk assessment databases for Missouri. The risk areas assessed included individual susceptibility or risk, potential transmission or community risk, socioeconomic contextual risk, accessibility constraints, health culture risk, and, finally, the exposure risk based on current case loads of COVID-19 at the county level. The goal of this project is to support data-driven decision-making processes across levels of government and health care providers to enable incorporation of significant risk factors associated with their specific populations and potential synergies and enable preparation for resilience and mitigation efforts across rural counties.

The GeoARK data extraction and build for these risk databases included a selection of 325 (5.91%) elements from the current catalogue of 5500, integrating 35 different sources. A subset (total of 91 [28%] across all 6 areas) were then selected for use in the calculation of total risk scores for each assessment area. More specifically, components included known and possible comorbidities and age breaks; commuting, migration, worker types, group gatherings, and living situation; race, ethnicity, disability, insurance status, veteran status, and education level; development and inclusion of various hospital, nursing homes, and telehealth access measures; and broadband metrics. Ordinary least squares regression was used to evaluate combinations of explanatory variables. Selected variables within each risk category then had quintiles calculated to create comparative categorical groups for each risk variable, with higher values assigned to worse risk. Cumulative risk scores were assembled for each risk category, as well as an overall composite risk score. These values were then analyzed using Local Moran's I, similarity analysis, and spatially constrained multivariate clustering to inform regional grouping outcomes. Through spatial analytics, differences in both the magnitude of risk and the substance of that risk, among and between rural and urban counties, were found. Missouri's spatial diversity is evident in the variability of overall risk across the 6 factor areas developed as well as the 6 region-based groups of counties sharing similar risk traits. The results are queryable through the Geo-Context and COVID-19 website (Figure 4) [59].

These research results enhance the understanding of COVID-19 behavior and enable preparation for resilience in rural populations. It is important to understand the context and interrelationships of various risk factors occurring within the state in order to better understand the potential pathways for disease as well as what nuances in mitigation strategies are needed to address specific populations. There is no 1-size-fits-all solution for the diversity found through spatial analysis of risk. The ability to address issues that are most influencing the health of a particular region or population is paramount to equality in care.

Figure 4. Screen capture of the "Geo-Context and COVID-19 in Missouri" dashboard interface populated with GeoARK parameters. GeoARK: Geospatial Analytical Research Knowledgebase.



Discussion

Principal Findings

There are unique innovations interwoven within the design of the GeoARK system. It has taken the multilayered world of typical GIS analysis and flattened it. The incorporation at each location of keys (ie, geographic-level Federal Information Processing Standard [FIPS] code, zip code) creates bridges for associated attribution to be incorporated into the GeoARK-BT. Other information is integrated through geospatial location, leveraging the fact that the information occupies the same location on the earth's surface. For each point, the various scales, resolutions, information, and accuracies are captured as associated attributes of the particular data ingested. Through the integrated data services and analytical tools of this project, complex queries can be posed and associations explored. This is enabled only when spatial contexts have been quantified and thousands of factors associated spatially.

The enhanced analytics can provide a catalyst for health researchers to move beyond basic thematic mapping. In many, if not most, cases, the true benefit of using location is in the creation of new associations between data elements and subsequent creation of new information. The generation of this new quantified, tabular information is the real power of geospatial information and GeoARK.

Benefits and Opportunities

The GeoARK system facilitates the use and integration of geoinformatics within the broad health-based user community. There is a high level of effort and expertise required to locate, compile, transform, standardize, and then integrate the multifaceted data required to adequately understand the context of eHealth events. It is critically important that health research be buoyed with access to the GeoARK system as it decreases duplication of effort, allows comparisons across a much broader set of potential variables, and extends the breadth and scope of investigations beyond the boundaries of conventional variable thematic mapping.

The linkage of results to a specific geographic scale, and the concurrent interpretation of them in context, is a growing requirement of sociological and health research. Because the GeoARK system has precalculated and captured the distance-based relationships of neighbors, features, and other spatial context, the project will aid researchers in development of comparable populations at varying scales. This could be within a certain aspect of interest (rural-urban) or geography (county, zip code, or tract).

A focus of potential benefit will be the use of the GeoARK system in research design. Meaningful health analytics typically address developing and testing hypotheses to contrast and compare 1 group (reference) to another (comparison). The ability to "know" and possibly choose to control for "outside" variables (eg, environmental, social, cultural, infrastructure, or other factors) during the design of a study or trial may provide a clearer picture of the health aspect under investigation. The ability to tighten the research question or clinical trial, and its reference groups, leads to higher potential to achieve significant insights.

Challenges and Limitations

The patient protections provided through the implementation and interpretation of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) impacts the geographic scales at which we can investigate the detailed distributions of

disease and health effects. Although all disease occurs as events, the way in which aggregation, compilation, and subsequent roll-up of these events into geographies has a dampening effect on most, if not all, attempts to drill deeper into the spatial context and phenomenology of diseases.

Variability and uncertainty exist within all data collected by organizations as it was in the pursuit of a mandated purpose. Biases, ethical issues, and errors complicate the systematic integration of heterogeneous information into any database. By using location, it is hoped that these biases and other issues will be more clearly brought to light.

The modifiable areal unit problem has the potential to create problems with representation of certain types of data. By assembling these data across a variety of raster resolutions, the scales of representation can be tested and understood so that use of these data at any scale would be accompanied by a "fitness of use" measure that can be presented to the user. Because a range of geographies is captured upon integration within the design of GeoARK, these comparisons can also be tested for stability and significance across a range of scales. This allows researchers to evaluate at what level the component of interest manifests itself and therefore permit identification of the proper level for intervention (as well as what determinants are amenable to this process) or information that can be used for avoidance of a particular type of disparity in a particular area.

The evolution of the GeoARK-BT to the fully envisioned system as a web-based portal with robust data and research services has many hurdles to overcome. These include data usage agreements, compute scaling, cloud service strategy, data-as-a-service management strategy, security and compliance adjustments, performance tuning, build out of analytics, and cost constraints.

However, the biggest challenge remaining for health researchers is to learn to think spatially about their problems and broaden their research questions into the multifactor, multiscale arenas of investigation that the GeoARK system supports.

Conclusion

This paper describes and outlines the design, compilation, and assembly of the GeoARK system, a spatially referenced data table that facilitates the integration and standardization of sociocultural, infrastructural, environmental, and health-related data into a common, extractable, and analytical framework.

The GeoARK system provides the ability to identify, mitigate, and contextualize health disparities. It provides health researchers with an integrated big data repository that can be searched to enable stronger research designs, for example, develop sampling/surveillance approaches or clinical trial focus. Using context across a broad range of data, research topics surrounding avoidance, fairness, equity, justice, and acceptability within, or for, a given location can be pursued. GeoARK supports user-based query, contextual analysis, and visualization to investigate relationships among the integrated data layers as well as discover patterns of interest for health research.

There are myriad ways that the GeoARK system, as a service, can be used in future analyses in order to better understand health disparities and other research issues. This system enables researchers to draw deeper and more broadly applicable empirical evidence for health research and associated outcomes, as well as supporting AI/ML-friendly data extracts that can then leverage new spatial associations.

This framework provides benefit to eHealth-related research, applications, and policy evaluation by the broader health community and has the potential to transform health research from a layer-based mentality to an interactive integrated contextual knowledge platform.

Acknowledgments

The authors thank Katrina Boles, Kao Yang, Sam Spell, and Rebecca Shyu at the University of Missouri for their support and assistance with initial development phases. We also acknowledge our case study collaborators, Dr Saif Khairat, Dr Eileen Avery, and Dr Richard Hammer, for their research utilizing the Geospatial Analytical Research Knowledgebase (GeoARK) system and providing feedback on the development and usability of the system.

TH was supported by the National Institutes of Health (NIH-5T32LM012410). CRS was supported by the National Science Foundation (NSF) Division Information and Intelligent Systems (Award IIS-2027891) and the University of Missouri System Research and Creative Works Strategic Investment Program. The cyber infrastructure hosting the geospatial big table was supported in part by the NSF Computer and Network Systems (CNS-1429294). This paper's content is solely the responsibility of the authors and does not represent the official views of the NIH or the NSF.

Data Availability

The data underlying this paper will be shared through the Geospatial Analytical Research Knowledgebase (GeoARK) service. Questions can be sent to the corresponding author.

Authors' Contributions

TH and CRS designed the Geospatial Analytical Research Knowledgebase (GeoARK) system and contributed to the design and conduct of the case studies. TH led the writing of the manuscript. DL implemented and indexed NoSQL databases, performed

data loads, and assessed query efficiencies. TY provided support and assistance with metadata processes and web portal interface development. CRS oversaw the project and obtained research funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary File providing 3 pages of both the tabular data source listing and the geospatial source listing of files integrated or being integrated into the GeoARK-BT and available through the GeoARK system for support of complex queries and data selection in support of the health research community. GeoARK: Geospatial Analytical Research Knowledgebase; GeoARK-BT: GeoARK-Big Table.
[PDF File (Adobe PDF File), 841 KB - medinform_v10i4e35073_app1.pdf]

References

1. Sun W, Gong F, Xu J. Individual and contextual correlates of cardiovascular diseases among adults in the United States: a geospatial and multilevel analysis. GeoJournal 2019 Jul 25;85(6):1685-1700. [doi: 10.1007/s10708-019-10049-7]
2. Robles B, Thomas CS, Lai ES, Kuo T. A geospatial analysis of health, mental health, and stressful community contexts in Los Angeles County. Prev Chronic Dis 2019 Nov 07;16:E150 [FREE Full text] [doi: 10.5888/pcd16.190138] [Medline: 31701869]
3. Graham H. Social determinants and their unequal distribution: clarifying policy understandings. Milbank Q 2004 Mar;82(1):101-124 [FREE Full text] [doi: 10.1111/j.0887-378x.2004.00303.x] [Medline: 15016245]
4. Richardson DB, Volkow ND, Kwan M, Kaplan RM, Goodchild MF, Croyle RT. Medicine. Spatial turn in health research. Science 2013 Mar 22;339(6126):1390-1392 [FREE Full text] [doi: 10.1126/science.1232257] [Medline: 23520099]
5. Wong MS, Ho HC, Tse A. Geospatial context of social and environmental factors associated with health risk during temperature extremes: review and discussion. Geospat Health 2020 Jun 22;15(1):168-173 [FREE Full text] [doi: 10.4081/gh.2020.814] [Medline: 32575974]
6. Phillips IG, McCuskey DJ, Felt D, Raman AB, Hayford CS, Pickett J, et al. Geospatial perspectives on health: the PrEP4Love campaign and the role of local context in health promotion messaging. Soc Sci Med 2020 Nov;265:113497 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113497] [Medline: 33187750]
7. Eberhardt MS, Pamuk ER. The importance of place of residence: examining health in rural and nonrural areas. Am J Public Health 2004 Oct;94(10):1682-1686. [doi: 10.2105/ajph.94.10.1682] [Medline: 15451731]
8. LaVeist T, Pollack K, Thorpe R, Fesahazion R, Gaskin D. Place, not race: disparities dissipate in southwest Baltimore when blacks and whites live under similar conditions. Health Aff (Millwood) 2011 Oct;30(10):1880-1887 [FREE Full text] [doi: 10.1377/hlthaff.2011.0640] [Medline: 21976330]
9. Arcaya MC, Tucker-Seeley RD, Kim R, Schnake-Mahl A, So M, Subramanian S. Research on neighborhood effects on health in the United States: a systematic review of study characteristics. Soc Sci Med 2016 Nov;168:16-29 [FREE Full text] [doi: 10.1016/j.socscimed.2016.08.047] [Medline: 27637089]
10. Scribner RA, Simonsen NR, Leonardi C. The social determinants of health core: taking a place-based approach. Am J Prev Med 2017 Jan;52(1S1):S13-S19 [FREE Full text] [doi: 10.1016/j.amepre.2016.09.025] [Medline: 27989288]
11. Estima J, Painho M. User generated spatial content-integrator: conceptual model to integrate data from diverse sources of user generated spatial content. IJGI 2016 Oct 09;5(10):183. [doi: 10.3390/ijgi5100183]
12. Wong ST, Regan S. Patient perspectives on primary health care in rural communities: effects of geography on access, continuity and efficiency. RRH 2009 Mar 18;9(1):1142. [doi: 10.22605/rrh1142]
13. Bissonnette L, Wilson K, Bell S, Shah TI. Neighbourhoods and potential access to health care: the role of spatial and aspatial factors. Health Place 2012 Jul;18(4):841-853 [FREE Full text] [doi: 10.1016/j.healthplace.2012.03.007] [Medline: 22503565]
14. Stingone JA, Buck Louis GM, Nakayama SF, Vermeulen RC, Kwok RK, Cui Y, et al. Toward greater implementation of the exposome research paradigm within environmental epidemiology. Annu Rev Public Health 2017 Mar 20;38(1):315-327 [FREE Full text] [doi: 10.1146/annurev-publhealth-082516-012750] [Medline: 28125387]
15. Guthman J, Mansfield B. The implications of environmental epigenetics: a new direction for geographic inquiry on health, space, and nature-society relations. Prog Hum Geogr 2012 Nov 26;37(4):486-504. [doi: 10.1177/0309132512463258]
16. Keet CA, McCormack MC, Pollack CE, Peng RD, McGowan E, Matsui EC. Neighborhood poverty, urban residence, race/ethnicity, and asthma: rethinking the inner-city asthma epidemic. J Allergy Clin Immunol 2015 Mar;135(3):655-662 [FREE Full text] [doi: 10.1016/j.jaci.2014.11.022] [Medline: 25617226]
17. Greenough PG, Nelson EL. Beyond mapping: a case for geospatial analytics in humanitarian health. Confl Health 2019 Nov 08;13(1):50 [FREE Full text] [doi: 10.1186/s13031-019-0234-9] [Medline: 31719842]
18. Towne SD. Socioeconomic, geospatial, and geopolitical disparities in access to health care in the US 2011-2015. Int J Environ Res Public Health 2017 May 29;14(6):573 [FREE Full text] [doi: 10.3390/ijerph14060573] [Medline: 28555045]

19. Downey LH. Rural populations and health: determinants, disparities, and solutions. Prev Chronic Dis 2013 Jun 27;10:E104. [doi: 10.5888/pcd10.130097]
20. Marcin JP, Shaikh U, Steinhorn RH. Addressing health disparities in rural communities using telehealth. Pediatr Res 2016 Jan 14;79(1-2):169-176 [FREE Full text] [doi: 10.1038/pr.2015.192] [Medline: 26466080]
21. Cannuscio CC, Weiss EE, Asch DA. The contribution of urban foodways to health disparities. J Urban Health 2010 May 31;87(3):381-393 [FREE Full text] [doi: 10.1007/s11524-010-9441-9] [Medline: 20354910]
22. Turner N, Pan W, Martinez-Bianchi V, Panayotti G, Planey A, Woods C, et al. Racial, ethnic, and geographic disparities in novel coronavirus (severe acute respiratory syndrome coronavirus 2) test positivity in North Carolina. Open Forum Infect Dis 2021 Jan;8(1):ofaa413 [FREE Full text] [doi: 10.1093/ofid/ofaa413] [Medline: 33575416]
23. Shakoor H, Jehan N, Khan S, Khattak NU. Investigation of radon sources, health hazard and risks assessment for children using analytical and geospatial techniques in District Bannu (Pakistan). Int J Radiat Biol 2021 Mar 15:1-9. [doi: 10.1080/09553002.2021.1872817] [Medline: 33428859]
24. Long X, Liu F, Zhou X, Pi J, Yin W, Li F, et al. Estimation of spatial distribution and health risk by arsenic and heavy metals in shallow groundwater around Dongting Lake plain using GIS mapping. Chemosphere 2021 May;269:128698. [doi: 10.1016/j.chemosphere.2020.128698] [Medline: 33121802]
25. Adimalla N, Qian H. Geospatial distribution and potential noncarcinogenic health risk assessment of nitrate contaminated groundwater in southern India: a case study. Arch Environ Contam Toxicol 2020 Oct 03;80(1):107-119. [doi: 10.1007/s00244-020-00762-7]
26. Khairat S, Haithcoat T, Liu S, Zaman T, Edson B, Gianforcaro R, et al. Advancing health equity and access using telemedicine: a geospatial assessment. J Am Med Inform Assoc 2019 Aug 01;26(8-9):796-805 [FREE Full text] [doi: 10.1093/jamia/ocz108] [Medline: 31340022]
27. Downey L, Van Willigen M. Environmental stressors: the mental health impacts of living near industrial activity. J Health Soc Behav 2005 Oct 24;46(3):289-305 [FREE Full text] [doi: 10.1177/002214650504600306] [Medline: 16259150]
28. Shrestha R, Flacke J, Martinez J, van Maarseveen M. Environmental health related socio-spatial inequalities: identifying "hotspots" of environmental burdens and social vulnerability. Int J Environ Res Public Health 2016 Jul 09;13(7):691 [FREE Full text] [doi: 10.3390/ijerph13070691] [Medline: 27409625]
29. Wilhelm M, Qian L, Ritz B. Outdoor air pollution, family and neighborhood environment, and asthma in LA FANS children. Health Place 2009 Mar;15(1):25-36 [FREE Full text] [doi: 10.1016/j.healthplace.2008.02.002] [Medline: 18373944]
30. Jerrett M, Burnett RT, Beckerman BS, Turner MC, Krewski D, Thurston G, et al. Spatial analysis of air pollution and mortality in California. Am J Respir Crit Care Med 2013 Sep;188(5):593-599. [doi: 10.1164/rccm.201303-0609oc]
31. Ruiz MO, Tedesco C, McTighe TJ, Austin C, Kitron U. Environmental and social determinants of human risk during a West Nile virus outbreak in the Greater Chicago area. Int J Health Geogr 2002;3(1):8. [doi: 10.1186/1476-072x-3-8]
32. Pearce J, Richardson E, Mitchell R, Shortt N. Environmental justice and health: the implications of the socio-spatial distribution of multiple environmental deprivation for health inequalities in the United Kingdom. Trans Inst Br Geogr 2010;35(4):522-539. [doi: 10.1111/j.1475-5661.2010.00399.x]
33. Pearce J, Cherrie M, Shortt N, Deary I, Ward Thompson C. Life course of place: a longitudinal study of mental health and place. Trans Inst Br Geogr 2018 May 02;43(4):555-572. [doi: 10.1111/tran.12246]
34. Shortt NK, Rind E, Pearce J, Mitchell R, Curtis S. Alcohol risk environments, vulnerability and social inequalities in alcohol consumption. Ann Am Assoc Geogr 2018 Mar 21;108(5):1210-1227 [FREE Full text] [doi: 10.1080/24694452.2018.1431105] [Medline: 32154488]
35. Johnston R, Poulsen M, Forrest J. Research note—measuring ethnic residential segregation: putting some more geography in. Urban Geogr 2013 May 16;30(1):91-109. [doi: 10.2747/0272-3638.30.1.91]
36. Bull CN, Krout JA, Rathbone-McCuan E, Shreffler MJ. Access and issues of equity in remote/rural areas. J Rural Health 2001 Sep;17(4):356-359. [doi: 10.1111/j.1748-0361.2001.tb00288.x] [Medline: 12071561]
37. Luo W, Qi Y. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. Health Place 2009 Dec;15(4):1100-1107. [doi: 10.1016/j.healthplace.2009.06.002] [Medline: 19576837]
38. Amstislavski P, Matthews A, Sheffield S, Maroko AR, Weedon J. Medication deserts: survey of neighborhood disparities in availability of prescription medications. Int J Health Geogr 2012;11(1):48. [doi: 10.1186/1476-072x-11-48]
39. Chakraborty J, Maantay J. Proximity analysis for exposure assessment in environmental health justice research. In: Maantay JA, McLafferty S, editors. Geospatial Analysis of Environmental Health. The Netherlands: Springer; 2011:111-138.
40. Lynch SM, Wiese D, Ortiz A, Sorice KA, Nguyen M, González ET, et al. Towards precision public health: geospatial analytics and sensitivity/specificity assessments to inform liver cancer prevention. SSM Popul Health 2020 Dec;12:100640 [FREE Full text] [doi: 10.1016/j.ssmph.2020.100640] [Medline: 32885020]
41. Gilliland JA, Shah TI, Clark A, Sibbald S, Seabrook JA. A geospatial approach to understanding inequalities in accessibility to primary care among vulnerable populations. PLoS One 2019 Jan 7;14(1):e0210113 [FREE Full text] [doi: 10.1371/journal.pone.0210113] [Medline: 30615678]
42. Planey AM. Audiologist availability and supply in the United States: a multi-scale spatial and political economic analysis. Soc Sci Med 2019 Feb;222:216-224. [doi: 10.1016/j.socscimed.2019.01.015] [Medline: 30660682]

43. Biggeri A, Barbone F, Lagazio C, Bovenzi M, Stanta G. Air pollution and lung cancer in Trieste, Italy: spatial analysis of risk as a function of distance from sources. Environ Health Perspect 1996 Jul;104(7):750-754 [FREE Full text] [doi: 10.1289/ehp.96104750] [Medline: 8841761]
44. Schuurman N, Peters PA, Oliver LN. Are obesity and physical activity clustered? A spatial analysis linked to residential density. Obesity (Silver Spring) 2009 Dec;17(12):2202-2209 [FREE Full text] [doi: 10.1038/oby.2009.119] [Medline: 19390521]
45. Chandra S, Kassens-Noor E, Kuljanin G, Vertalka J. A geographic analysis of population density thresholds in the influenza pandemic of 1918–19. Int J Health Geogr 2013;12(1):9. [doi: 10.1186/1476-072x-12-9]
46. Mizen A, Fry R, Rodgers S. GIS-modelled built-environment exposures reflecting daily mobility for applications in child health research. Int J Health Geogr 2020 Apr 10;19(1):12 [FREE Full text] [doi: 10.1186/s12942-020-00208-2] [Medline: 32276644]
47. Centers for Disease Control and Prevention. Interactive Atlas of Heart Disease and Stroke. URL: http://nccd.cdc.gov/DHDSPAtlas [accessed 2022-03-31]
48. United States Geological Survey. The National Map. URL: https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map [accessed 2022-03-31]
49. United States Department of Agriculture. Atlas of Rural and Small-Town America. URL: https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/ [accessed 2022-03-31]
50. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
51. Danko D. ISO/TC211: geographic information – metadata ISO 19115. In: Moellering H, Aalders HJGL, Crane A, editors. World Spatial Metadata Standards. Oxford: Elsevier Science; 2005:535-555.
52. Bambra C, Gibson M, Sowden A, Wright K, Whitehead M, Petticrew M. Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. J Epidemiol Community Health 2010 May 19;64(4):284-291 [FREE Full text] [doi: 10.1136/jech.2008.082743] [Medline: 19692738]
53. Egerter S, Barclay C, Grossman-Kahn R, Braveman P. Exploring the social determinants of health. Violence, Social Disadvantage and Health, Robert Wood Johnson Foundation. In: The Foundation for Vulnerable Population's. NJ: Princeton; 2011:1-9.
54. Kyle AD, Woodruff TJ, Buffler PA, Davis DL. Use of an index to reflect the aggregate burden of long-term exposure to criteria air pollutants in the United States. Environ Health Perspect 2002 Mar;110 Suppl 1(suppl 1):95-102 [FREE Full text] [doi: 10.1289/ehp.02110s195] [Medline: 11834467]
55. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System Survey Data. URL: https://www.cdc.gov/brfss/data_documentation/index.htm [accessed 2022-03-30]
56. United States Census Bureau. Explore Census Data. URL: https://data.census.gov/cedsci/ [accessed 2022-03-31]
57. Anselin L, Sridharan S, Gholston S. Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. Soc Indic Res 2006 Nov 15;82(2):287-309. [doi: 10.1007/s11205-006-9034-x]
58. Haithcoat TL, Avery EE, Bowers KA, Hammer RD, Shyu C. Income inequality and health: expanding our understanding of state-level effects by using a geospatial big data approach. Soc Sci Comput Rev 2019 Sep 03;39(4):543-561. [doi: 10.1177/0894439319872991]
59. University of Missouri, IDAS Lab. Geo-Context and COVID-19 in Missouri. URL: https://geoark.missouri.edu/counties [accessed 2022-03-31]

Abbreviations

ACS: American Community Survey
AI: artificial intelligence
FIPS: Federal Information Processing Standard
GeoARK: Geospatial Analytical Research Knowledgebase
GeoARK-BT: GeoARK-Big Table
GIS: Geographic Information System
ML: machine learning
NIH: National Institutes of Health

Edited by C Lovis; submitted 19.11.21; peer-reviewed by S Pesälä, SY Shin; comments to author 25.02.22; revised version received 27.02.22; accepted 11.03.22; published 06.04.22.

Please cite as:
Haithcoat T, Liu D, Young T, Shyu CR
Investigating Health Context Using a Spatial Data Analytical Tool: Development of a Geospatial Big Data Ecosystem
JMIR Med Inform 2022;10(4):e35073
URL: https://medinform.jmir.org/2022/4/e35073
doi:10.2196/35073
PMID:35311683

Original Paper

Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study

Khaled El Emam[1,2,3], BEng, PhD; Lucy Mosquera[2,3], BA, MSc; Xi Fang[3], BA, MSc; Alaa El-Hussuna[4], MSc, MD

[1]School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

[2]Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada

[3]Replica Analytics Ltd, Ottawa, ON, Canada

[4]Open Source Research Collaboration, Aarlberg, Denmark

Corresponding Author:
Khaled El Emam, BEng, PhD
School of Epidemiology and Public Health
University of Ottawa
401 Smyth Road
Ottawa, ON, K1H 8L1
Canada
Phone: 1 6137975412
Email: kelemam@ehealthinformation.ca

Abstract

Background: A regular task by developers and users of synthetic data generation (SDG) methods is to evaluate and compare the utility of these methods. Multiple utility metrics have been proposed and used to evaluate synthetic data. However, they have not been validated in general or for comparing SDG methods.

Objective: This study evaluates the ability of common utility metrics to rank SDG methods according to performance on a specific analytic workload. The workload of interest is the use of synthetic data for logistic regression prediction models, which is a very frequent workload in health research.

Methods: We evaluated 6 utility metrics on 30 different health data sets and 3 different SDG methods (a Bayesian network, a Generative Adversarial Network, and sequential tree synthesis). These metrics were computed by averaging across 20 synthetic data sets from the same generative model. The metrics were then tested on their ability to rank the SDG methods based on prediction performance. Prediction performance was defined as the difference between each of the area under the receiver operating characteristic curve and area under the precision-recall curve values on synthetic data logistic regression prediction models versus real data models.

Results: The utility metric best able to rank SDG methods was the multivariate Hellinger distance based on a Gaussian copula representation of real and synthetic joint distributions.

Conclusions: This study has validated a generative model utility metric, the multivariate Hellinger distance, which can be used to reliably rank competing SDG methods on the same data set. The Hellinger distance metric can be used to evaluate and compare alternate SDG methods.

(JMIR Med Inform 2022;10(4):e35734) doi:10.2196/35734

KEYWORDS

synthetic data; data utility; data privacy; generative models; utility metric; synthetic data generation; logistic regression; model validation; medical informatics; binary prediction model; prediction model

Introduction

Interest in synthetic data generation (SDG) has recently grown. Synthetic data are deemed to have low privacy risks in practice because there is no one-to-one mapping between synthetic records and real people [1-8]. Recent evidence supports the low privacy risk claim [9]. This enables synthetic data to be used and shared for secondary purposes without the need for further consent [10]. In addition to meeting privacy requirements, synthetic data must also have sufficient utility. This utility can be evaluated using utility metrics. Utility metrics are important in hyperparameter tuning of the generative models during

training and communicating data quality to the synthetic data users and for researchers and analysts when ranking different SDG methods to select the best one. Our focus in this paper is on the ranking of SDG methods.

Utility metrics can be defined as narrow or broad [11]. Narrow metrics are specific to an analysis that is performed with the synthetic data and are also sometimes referred to as workload-aware utility metrics. For example, if the objective is to build a model between a predictor and a binary outcome, controlling for multiple confounders, then the difference in accuracy of real versus synthetic model predictions on holdout data sets would be a workload-aware utility metric. There have been multiple studies evaluating narrow metrics [12-16]. Narrow metrics represent what the data user is ultimately interested in. Data users want synthetic data sets that score highly on narrow utility metrics.

Researchers and analysts need to rank SDG methods. For example, a developer of an SDG method may use an ensemble of techniques and then select the one with the highest utility as the final result, or analysts may evaluate multiple SDG methods available in the marketplace to select one for their own projects. However, all workloads are typically not known in advance. Therefore, researchers and analysts cannot evaluate the narrow utility of the SDG methods directly. Instead, they need to use broad utility metrics during the SDG construction and evaluation process. A key requirement is that broad utility metrics are predictive of narrow utility metrics for plausible analytic workloads.

Some studies utilized broad metrics, for example, to compare and improve SDG methods [17-19]. However, many of the broad utility metrics currently used have not been validated. This means that there is a dearth of evidence demonstrating that they are predictive of narrow utility metrics under realistic decision-making scenarios.

The realistic decision-making scenario that we are considering here is the comparison and ranking of SDG methods. Finding the best SDG method is becoming a more common need in the literature, and we need reliable metrics to be able to draw valid conclusions from these comparisons. Furthermore, in practice, users of SDG methods need to have good metrics to select among a number of these methods that may be available to them.

Utility metrics can be classified in a different way, which is relevant for our purposes. They can pertain to a specific synthetic data set or to the generative model ("data set–specific" and "model-specific" utility metrics). Because SDG is stochastic, the utility of synthetic data sets generated from the same generative model will vary each time the generative model is run, and sometimes that variation can be substantial. Data set–specific utility metrics are useful when one wants to communicate how good the particular generated data set is to a data user. However, these utility metrics are not necessarily useful, for example, for comparing different generative models because of the stochasticity. A model-specific utility metric reflects the utility of the generated synthetic data sets on average, across many data sets that are generated from the same model. Such a metric is more useful in our context, where we want to compare and rank SDG methods.

Our focus in the current study is to perform a validation study of broad model-specific utility metrics for structured (tabular) health data sets. While there have been evaluations of generative model utility metrics in the past, these have focused on images rather than structured data [20]. One previous more relevant evaluation considered propensity mean squared error (pMSE) [21,22] as a model utility metric whereby its correlation with binary prediction accuracy on synthetic data was empirically assessed [23]. The authors found that when used as a broad model-specific utility metric, by averaging across multiple synthetic data sets, this metric had a moderate correlation with narrow model-specific utility metrics. However, the correlation between a broad metric and a narrow metric across many data sets for a single SDG technique does not reflect an actual decision-making scenario. In practice, we have a single data set and multiple SDG techniques. Therefore, the extent to which the results from that previous study would be informative to our scenario of interest is unclear.

We build on this previous work by considering other types of broad model-specific utility metrics beyond pMSE and adjust the methodology to more closely model a practical decision-making scenario of an analyst selecting among multiple SDG methods to identify the one with higher narrow utility on logistic regression prediction tasks. This type of prediction task is used often in health research.

Methods

The protocol for this study was approved by the CHEO Research Institute Research Ethics Board (number CHEOREB# 21/144X). Our objective was to answer the following question: Which broad model-specific utility metrics can be used to rank SDG methods in terms of the similarity of prediction performance between real and synthetic data? In the following sections we describe the methods that were followed.

Data Sets

For our analysis, we used the 30 health data sets that are summarized in Appendix S1 in Multimedia Appendix 1. These data sets are available publicly or can be requested from the data custodians. Many of these data sets have been used in previous evaluations of SDG techniques [12,15,23], and therefore we can ensure some consistency across studies in this domain. These data sets also represent a heterogeneous set of clinical situations (providing care, observational studies, clinical trials, and registries), a wide range of data set sizes (87-44,842 patients), and variation in data set complexity (as measured using average variable entropy), which allow our evaluations to be more generalizable.

The Broad Utility Metrics Considered

Broad utility metrics compare the joint distributions of the real and synthetic data sets. Many metrics have been proposed to compare joint distributions [24]. We only focus on 6 multivariate metrics that have been used in previous work to evaluate the utility of synthetic data sets.

Maximum Mean Discrepancy

The maximum mean discrepancy metric is one way to test whether samples are from different distributions [25]. In our implementation, we used a radial basis function kernel. This metric has been applied to assess the utility of synthetic health data [26,27]. It is also widely used in the training of deep learning models and evaluation of the quality of synthetic data. Recent work on a recurrent Generative Adversarial Network (GAN) and recurrent conditional GAN made use of maximum mean discrepancy to assess whether the time series generated by the generative model implicitly learns the distribution of the true data [28]. Another study evaluated synthetic data in the smart grid context, in which a GAN is used to learn the conditional probability distribution of the significant features in the real data set and generates synthetic data based on the learnt distribution [29].

Multivariate Hellinger Distance

The Hellinger distance [30] has been shown to behave in a consistent manner as other distribution comparison metrics, specifically in the context of evaluating disclosure control methods [31], when comparing original and transformed data.

The Hellinger distance can be derived from the multivariate normal Bhattacharyya distance and has the advantage that it is bound between 0 and 1 and hence is more interpretable [32]. We constructed Gaussian copulas from the original and synthetic data sets [33] and then computed the distance between them. The concept of comparing the distance between 2 multivariate Gaussian distributions has been used to train GAN-based SDG methods [34]. Additional details on its calculation are provided in Appendix S2 in Multimedia Appendix 1.

Wasserstein Distance

The W_1 Wasserstein distance [35] is often applied to the training of GANs [36]. It has resulted in a learning process that is more robust by alleviating the vanishing gradient issue and mode collapse.

While GANs have been used extensively as an SDG technique, they very often still have trouble capturing the temporal dependency of the joint probability distributions caused by time-series data. The conditional sig-Wasserstein GANs proposed for time series generation is aimed at addressing this problem [37]. Here, the authors combine the signature of paths, which statistically describe the stream of data, and the W_1 distance, to capture the joint law of time series. By employing the sig-W as the discriminator, sig-Wasserstein GAN shows an ability to generate realistic multidimensional time series. Additional details on its calculation are provided in Appendix S2 in Multimedia Appendix 1.

Cluster Analysis Measure

The original cluster metric [21] was first purposed as a global measure of the data utility of original data and masked data. The cluster analysis has 2 steps: first, merge the original data (O) and masked data (M); then, given a certain number of groups G, perform cluster analysis on the merged data. The measure can be calculated as:

Where, n_j denotes the number of observations in the jth cluster and n_{jo} denotes the number of observations in the jth cluster that are from the original data (O). The c value is defined as:

A large U_c value indicates the disparities of the underlying latent structure of the original and masked data. The weight w_j can reflect the importance of certain clusters. This cluster analysis measure is used in the evaluation of synthetic data by simply replacing the original data with real data and the masked data with synthetic data [17].

Distinguishability Metrics

These broad metrics are based on the idea of training a binary classifier that can discriminate between a real and synthetic record [38,39]. That ability to discriminate is converted into a score.

A propensity mean square error metric has been proposed to evaluate the similarity of real and synthetic data sets [21,22], a perspective adopted from the propensity score matching literature [40], which we will refer to as *propensityMSE*. To calculate the *propensityMSE*, a classifier is trained on a stacked data set consisting of real observations labelled 1 and synthetic observations labelled 0. The *propensityMSE* score is computed as the mean squared difference of the estimated probability from the average prediction where it is not possible to distinguish between the 2 data sets. If the data sets are of the same size, which is the assumption we make here, and indistinguishable, then the average estimate will be 0.5.

Another related approach that has been used to evaluate the utility of synthetic data is to take a prediction perspective rather than a propensity perspective. This has been applied with "human discriminators" by asking a domain expert to manually classify sample records as real or synthetic [41-43]. This means that a sample of real records and a sample of synthetic records are drawn, and the 2 sets are shuffled together. Then the shuffled records are presented to clinicians who are experts in the domain, and they are asked to subjectively discriminate between the records by indicating which is real versus synthetic. High distinguishability only occurs when the human discriminator can correctly classify real and synthetic records.

The use of human discriminators is not scalable and therefore we can use machine learning algorithms trained on a training data set and that make predictions on a holdout test data set. This approach mimics the subjective evaluations described above. We will refer to this metric as *predictionMSE*. Also note that this calculation is different from the calculation of *propensityMSE* where the training data set is also used to compute the probabilities. Additional details on the calculations are provided in Appendix S2 in Multimedia Appendix 1.

Workload Aware (Narrow) Metrics

To assess whether the utility metrics are useful, we evaluated whether they can accurately rank SDG methods on workload

aware metrics. This section describes these workload aware metrics.

We built a logistic regression (LR) model for each data set. LR is common in health research, and a recent systematic review has shown that its performance is comparable to that of machine learning models for clinical prediction workloads [44]. Furthermore, an evaluation of the relative accuracy of LR models compared to that of other machine learning techniques, such as random forests and support vector machines, on synthetic versus real data sets across multiple types of SDG methods showed that LR models are only very slightly different [23]. Therefore, we would expect that the results using LR would provide broadly applicable and meaningful results.

We evaluated the prediction accuracy using 3-fold crossvalidation. Accuracy was measured using the area under the receiver operating characteristic curve (AUROC) [45] and the area under the precision-recall curve (AUPRC) [46]. For outcomes that had multiple categories, we used the average of pairwise AUROC values [47]. The AUPRC values for multicategory outcomes were macroaveraged. This was performed for each real and each synthetic data set.

To assess the similarity between the AUROC and AUPRC for the real and synthetic data sets, we computed the absolute difference between them. This provides a measure of how similar the real results are to the synthetic results.

Evaluation Methodology

For each of the 30 real data sets, we generated 20 synthetic data sets. The utility metrics and the absolute AUROC difference and absolute AUPRC difference were computed on each of the 20 synthetic data sets, and each of these was averaged. Therefore, for each of the data sets, we had 1 average utility metric value for each of the 6 utility metrics, 1 average AUROC difference value, and 1 average AUPRC difference value. These values are tabulated in Appendix S3 and S4 in Multimedia Appendix 1.

SDG Methods

The main hypothesis that we wanted to test was whether the utility metrics can be used to rank the SDG methods by their AUROC and AUPRC differences. The SDG methods were chosen to achieve representativeness, applicability, and variation.

1. Representativeness. The methods should reflect those that are often used in the community of practice and by researchers.
2. Applicability. The methods are those that an analyst would likely want to compare and select from to be consistent with our motivating use case.
3. Variation. The utility results among the chosen SDG methods should have variation sufficient for utility metrics to detect differences.

Three generative models were used: conditional GAN [48], a Bayesian network [49], and a sequential synthesis approach using decision trees [19]. The Bayesian network implementation uses a differential privacy approach. These 3 methods were selected for the following reasons: they each represent a class of methods that is often used in the literature (eg, sequential synthesis has been used on health and social sciences data [50-58], as well as Bayesian networks [26,59] and GANs [2,60,61]), they use very different approaches and therefore represent plausible SDG methods that an analyst would want to compare, and they are expected to exhibit large utility level variation given that different SDG methods tend to be better at modeling certain types of variables and relationships. For these 3 reasons, this set of SDG methods was suitable for this study on validating utility metrics.

Individual Utility Metric Ranking

We used the Page test to determine whether the utility metric prediction was correct [62]. For that, we specified 3 groups for each utility metric: an “L” group where the utility metric indicates low utility (ie, has the highest value since they are all distance-type metrics), an “H” group where the utility metric indicates high utility (ie, has the lowest value), and an “M” group in the middle. This process is repeated for each utility metric. For any particular data set, the generative model with the lowest utility is put in the “L” group, the generative model with the highest utility is put in the “H” group, and the third generative model is in the “M” group. Each generative model in a group is replaced with its AUROC or AUPRC difference value, depending on which workload aware metric is under evaluation.

The null hypotheses we were testing are therefore that:

$H0_{AUROC}$: median($AUROC\_Diff_L$) = median($AUROC\_Diff_M$) = median($AUROC\_Diff_H$)

$H0_{AUPRC}$: median($AUPRC\_Diff_L$) = median($AUPRC\_Diff_M$) = median($AUPRC\_Diff_H$)

where the subscript indicates the group. Against the alternatives:

$H1_{AUROC}$: median($AUROC\_Diff_L$) ≥ median($AUROC\_Diff_M$) ≥ median($AUROC\_Diff_H$)

$H1_{AUPRC}$: median($AUPRC\_Diff_L$) ≥ median($AUPRC\_Diff_M$) ≥ median($AUPRC\_Diff_H$)

Where at least one of the inequalities is strict. To compute the test statistic, *L*, the data are put in a matrix with 30 rows, one for each data set, and 3 columns, one for each group. The accuracy scores are used to assign a rank to the values in each row. Then the ranks are summed per column R_j where j=1…3. The *L* statistic is then the sum: $L = R_1 + 2R_2 + 3R_3$. The larger that value, the greater the evidence supporting the ranking conclusion.

Because of the relatively small sample size, we used an exact test of statistical significance. This also does not make distributional assumptions on the data, and for the number of data sets we have, this gives us a high-powered test.

If the test is significant, then the broad utility metric can be used to correctly rank SDG techniques based on their workload (narrow) metrics. Since we were comparing multiple utility metrics, a Bonferroni adjustment was made to the α level of .05 to account for multiple testing.

The maximum *L* value can be used to identify the utility metric that is best at ranking the SDG methods by prediction accuracy difference. This is particularly useful if more than one metric is found to be statistically significant.

Aggregate Ranking

Because each utility metric is expected to rank the SDG methods differently, we wanted to test whether an aggregate ranking would provide a better result than any of the individual utility metric rankings. We hoped to find an "ideal" ranking that has minimal distance to each of the individual rankings on the utility metrics. This can be performed for each data set separately, and then the ideal rankings across all the data sets would be evaluated on the Page test. The result would give us the performance of the aggregate ranking, and we can contrast that with the quality of individual utility metric rankings.

The distance we used is the Spearman footrule [63]. With this approach, if method A has a higher ranking than method B more often than not, method A should rank higher than method B in the ideal ranking. Given the relatively small data set, full enumeration rather than an optimization algorithm was used to find the ideal ranking.

Given that the *prediction*MSE and *propensity*MSE are strongly related, the former was removed so as to not give that particular ranking a higher weighting in the aggregation.

Results

The results of the ranking of the SDG methods are shown in Table 1. All metrics are statistically significant in that the null hypothesis of no difference was rejected. The broad utility metric rankings were close enough to the correct rank, so the relationship was quite strong.

The test statistic, the *L* value, indicates the strength of the ordering of data. The Hellinger distance had the highest *L* value among all the utility metrics, suggesting that it has an advantage in ordering the SDG methods based on their narrow utility metrics.

Table 1. Page test results for each of the utility metrics and prediction accuracy

Utility metric	AUROC[a] difference		AUPRC[b] difference	
	L value	*P* value	*L* value	*P* value
Maximum mean discrepancy	384	.00104[c]	392	<.001[c]
Hellinger distance[d]	398	<.001[c]	409	<.001[c]
Wasserstein distance	392	<.001[c]	403	<.001[c]
Cluster analysis	396,	<.001[c]	405	<.001[c]
Propensity mean squared error	390	<.001[c]	394	<.001[c]
Prediction mean squared error	396	<.001[c]	397	<.001[c]
Aggregate[d]	400	<.001[c]	408	<.001[c]

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

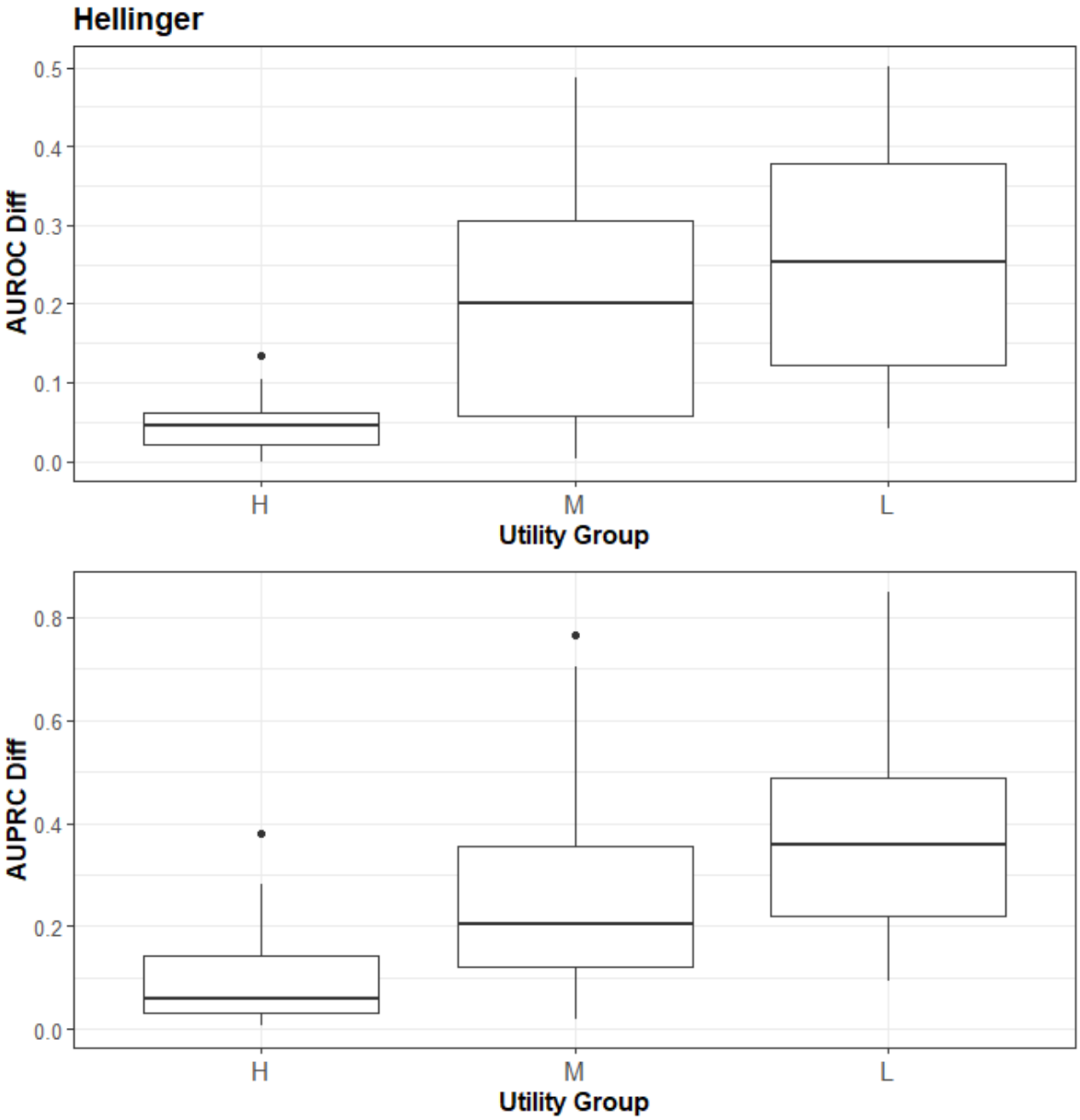
[b]AUPRC: area under the precision-recall curve.

[c]Statistically significant at a Bonferroni adjusted α level of .05.

[d]Highest metric on the test statistic.

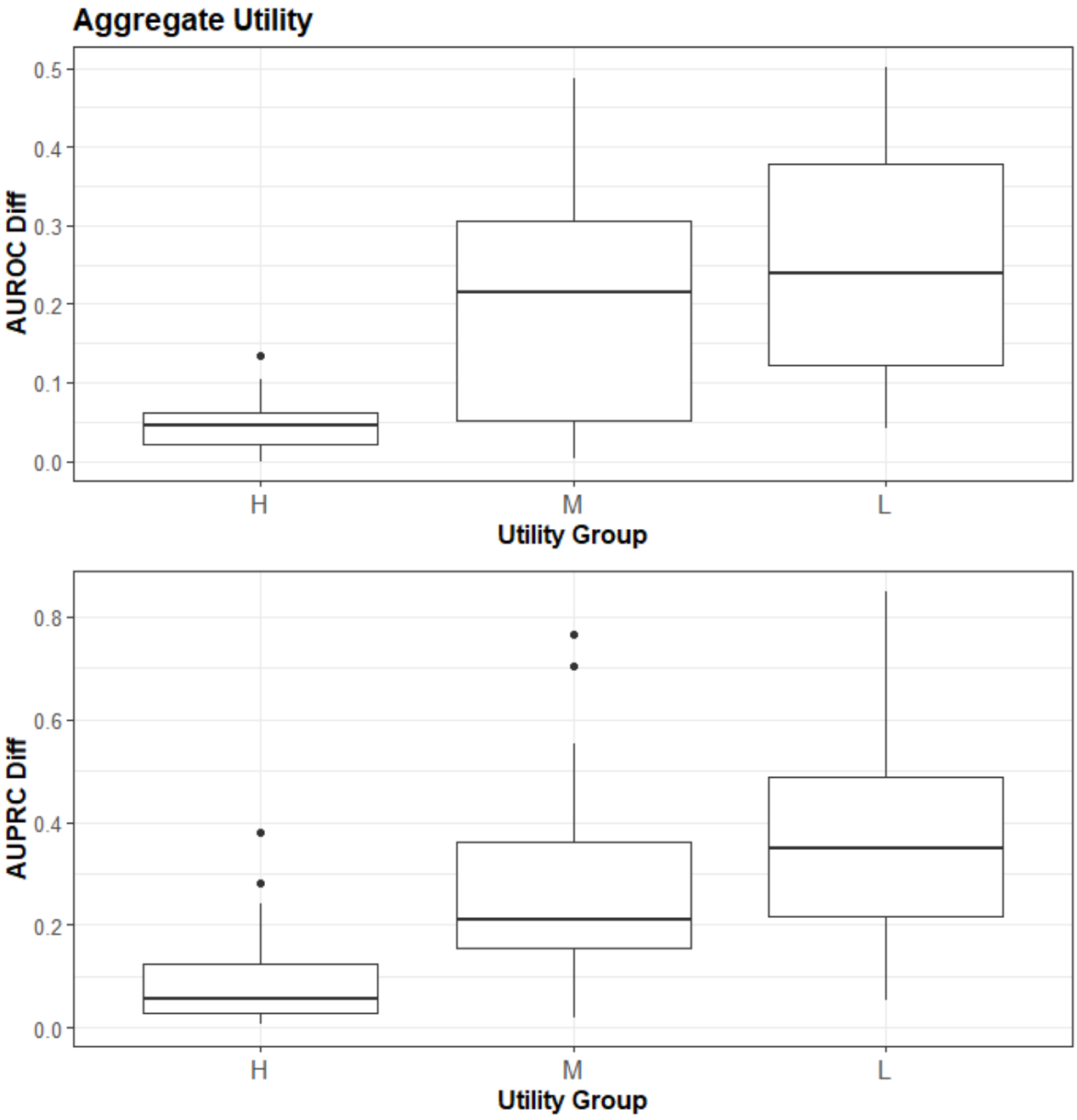
The boxplots in Figure 1 descriptively show the trend for the Hellinger distance. There is a clear trend of higher utility on the narrow AUROC and AUPRC metrics as the Hellinger distances get smaller. The boxplots for the remainder of the utility metrics are included in Appendix S5 in Multimedia Appendix 1, and they all show trends similar to those seen in Figure 1.

Figure 1. The relationship between the Hellinger distance versus the AUROC and AUPRC. The 3 SDG methods were ordered based on their relative Hellinger distance values into the "H," "M," and "L" groups. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; SDG: synthetic data generation.



The results for the aggregate ranking are shown in Table 1 and Figure 2. As can be seen from the *L* statistic and the boxplots, there is only a slight difference between using the Hellinger distance and the aggregate ranking from 5 different utility metrics. In a post-hoc analysis, we removed each of the metrics in turn in a leave-one-out fashion and recomputed the aggregate rank, but these did not produce better results than the one presented here.

Figure 2. The relationship between the aggregate ranking versus the AUROC and AUPRC. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve.



Discussion

Summary

The purpose of our study was to identify the most useful, broad generative model utility metrics. These are different from utility metrics calculated for a particular synthetic data set. Generative model utility characterizes the average utility across synthetic data sets that are produced from a generative model. Given the stochasticity of SDG, such utility metrics are more appropriate for evaluating, comparing, and selecting among SDG models on the same real data set. Single synthetic data set utility metrics, on the other hand, are useful for communicating synthetic data utility to a data user because these pertain to the particular synthetic data set that is being shared.

We performed our analysis using 3 types of generative models: a conditional GAN, a Bayesian network, and sequential decision trees. These 3 cover a broad cross-section of types of techniques that are used in practice, which would enhance the applicability and generalizability of the results.

In this study, we evaluated 6 different model-specific utility metrics to determine whether they can be used to rank SDG methods. This is a practical use case that reflects a decision that an analyst using SDG methods would want to make. For example, there are multiple SDG techniques that have been published in the literature, and our ranking results can help an analyst determine the one that would work best on their real data sets.

We defined workload-aware utility as the ability to develop binary or multicategory prediction models that have similar prediction accuracy, measured by the AUROC and the AUPRC, between the real and synthetic data sets. The construction of binary or multicategory prediction models is an often-used analytical workload for health data sets. We used logistic regression to compute the absolute AUROC and AUPRC differences on real and synthetic data sets.

Our results based on an evaluation on 30 heterogeneous health data sets indicated that all the utility metrics proposed in the literature will work well. However, the multivariate Hellinger distance computed over the Gaussian copula has a slight advantage in that it provides better utility ordering. Further examination of an aggregate ranking using multiple utility metrics showed only a negligible difference from the results of the Hellinger distance for the AUROC metric, and therefore the simplicity of a single utility metric would be preferred.

Our results would allow a researcher or analyst to select the SDG method with the highest utility defined in a narrow sense. However, maximum utility does not imply that the privacy risks are acceptably low. As there is a trade-off between utility and privacy, higher utility will increase the privacy risks as well. Therefore, when evaluating SDG methods, it is important to also consider the privacy risks.

Now that we have validation evidence for a broad utility metric, it can be combined with a privacy metric to provide an overall ranking of SDG methods. For example, membership disclosure metrics for generative models [64,65] can be considered along with the multivariate Hellinger distance when SDG methods are ranked. Metrics combining these 2 risk and utility metrics would be a good avenue for future research.

Limitations

An analyst may need to make other kinds of decisions, such as evaluating different SDG models for the purpose of hyperparameter tuning. Our study did not evaluate that specific use case, and therefore we cannot make broader claims that the Hellinger distance metric is suitable for other use cases.

Our study was performed by averaging the broad and narrow utility across 20 synthetic data sets (iterations). A larger number of iterations was evaluated (50 and 100), and we noted that the differences were not material. We opted to present the smaller number of iterations as these still give us meaningful results and would be faster computationally for others applying these results.

Acknowledgments

This study uses information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the website has contributed to or approved or is in any way responsible for the contents of this study. This research was enabled in part by support provided by Compute Ontario (computeontario.ca) and Compute Canada (www.computecanada.ca). This work was partially funded by the Canada Research Chairs program through the Canadian Institutes of Health Research, a Discovery Grant RGPIN-2016-06781 from the Natural Sciences and Engineering Research Council of Canada, through a contract with the Bill and Melinda Gates Foundation, and by Replica Analytics Ltd.

Conflicts of Interest

This work was performed in collaboration with Replica Analytics Ltd. This company is a spin-off from the Children's Hospital of Eastern Ontario Research Institute. KEE is co-founder and has equity in this company. LM and XF are data scientists employed by Replica Analytics Ltd.

Multimedia Appendix 1
Detailed SDG method descriptions, dataset descriptions, and detailed analysis results. SDG: synthetic data generation.
[PDF File (Adobe PDF File), 484 KB - medinform_v10i4e35734_app1.pdf]

References

1. Reiter JP. New approaches to data dissemination: a glimpse into the future (?). CHANCE 2012 Sep 20;17(3):11-15. [doi: 10.1080/09332480.2004.10554907]
2. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proc VLDB Endow 2018 Jun 01;11(10):1071-1083. [doi: 10.14778/3231751.3231757]
3. Hu J. Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data. arXiv. 2018. URL: http://arxiv.org/abs/1804.02784 [accessed 2022-03-01]
4. Taub J, Elliot M, Pampaka M, Smith D. Differential correct attribution probability for synthetic data: an exploration. In: Privacy in Statistical Databases. Switzerland: Springer, Cham; 2018:122-137.
5. Hu J, Reiter P, Wang Q. Disclosure risk evaluation for fully synthetic categorical data. In: Privacy in Statistical Databases. Switzerland: Springer, Cham; 2014:185-199.
6. Wei L, Reiter JP. Releasing synthetic magnitude microdata constrained to fixed marginal totals. SJI 2016 Feb 27;32(1):93-108. [doi: 10.3233/sji-160959]

7. Ruiz N, Muralidhar K, Domingo-Ferrer J. On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. In: Privacy in Statistical Databases. Switzerland: Springer, Cham; 2018:59-74.
8. Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. J Royal Statistical Soc A 2005 Jan;168(1):185-205. [doi: 10.1111/j.1467-985x.2004.00343.x]
9. El Emam K, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. J Med Internet Res 2020 Nov 16;22(11):e23139 [FREE Full text] [doi: 10.2196/23139] [Medline: 33196453]
10. El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation. Sebastopol, CA: O'Reilly Media, Inc; 2020.
11. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. Am Stat 2006 Aug;60(3):224-232. [doi: 10.1198/000313006x124640]
12. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open 2021 Apr 16;11(4):e043497 [FREE Full text] [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]
13. El Emam K, Mosquera L, Jonker E, Sood H. Evaluating the utility of synthetic COVID-19 case data. JAMIA Open 2021 Jan;4(1):ooab012 [FREE Full text] [doi: 10.1093/jamiaopen/ooab012] [Medline: 33709065]
14. Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, et al. Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. JMIR Med Inform 2020 Feb 20;8(2):e16492 [FREE Full text] [doi: 10.2196/16492] [Medline: 32130148]
15. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing. JMIR Med Inform 2020 Jul 20;8(7):e18910 [FREE Full text] [doi: 10.2196/18910] [Medline: 32501278]
16. Foraker RE, Yu SC, Gupta A, Michelson AP, Pineda Soto JA, Colvin R, et al. Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. JAMIA Open 2020 Dec;3(4):557-566 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa060] [Medline: 33623891]
17. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol 2020 May 07;20(1):108 [FREE Full text] [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]
18. Platzer M, Reutterer T. Holdout-Based Fidelity and Privacy Assessment of Mixed-Type Synthetic Data. arXiv. 2021 Apr 01. URL: http://arxiv.org/abs/2104.00635 [accessed 2022-10-01]
19. Emam KE, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. J Am Med Inform Assoc 2021 Jan 15;28(1):3-13 [FREE Full text] [doi: 10.1093/jamia/ocaa249] [Medline: 33186440]
20. Xu Q, Huang G, Yuan Y, Guo C, Sun Y, Wu F, et al. An empirical study on evaluation metrics of generative adversarial networks. arXiv. 2018. URL: http://arxiv.org/abs/1806.07755 [accessed 2022-11-01]
21. Woo MJ, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. JPC 2009 Apr 01;1(1):111-124 [FREE Full text] [doi: 10.29012/jpc.v1i1.568]
22. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. J R Stat Soc A 2018 Mar 07;181(3):663-688. [doi: 10.1111/rssa.12358]
23. Dankar FK, Ibrahim M. Fake it till you make it: guidelines for effective synthetic data generation. Appl Sci 2021 Feb 28;11(5):2158. [doi: 10.3390/app11052158]
24. Cha SH. Comprehensive survey on distance similarity measures between probability density functions. Math Models Methods Appl Sci 2007;4:300-307. [doi: 10.46300/9101]
25. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A. A Kernel Method for the Two-Sample Problem. In: Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. Cambridge, MA: MIT Press; 2007 Presented at: 20th Annual Conference on Neural Information Processing Systems: NIPS 200; December 4-7, 2006; Vancouver, BC URL: https://proceedings.neurips.cc/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf
26. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med 2020 Nov 09;3(1):147 [FREE Full text] [doi: 10.1038/s41746-020-00353-9] [Medline: 33299100]
27. Torfi A, Fox EA, Reddy CK. Differentially Private Synthetic Medical Data Generation using Convolutional GANs. arXiv. 2020. URL: http://arxiv.org/abs/2012.11774 [accessed 2022-11-01]
28. Cristóbal E, Stephanie L. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv. 2017. URL: https://arxiv.org/abs/1706.02633 [accessed 2021-11-01]
29. Zhang C, Kuppannagari SR, Kannan R, Prasanna VK. Generative Adversarial Network for Synthetic Time Series Data Generation in Smart Grids. 2018 Presented at: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm); October 29-31, 2018; Aalborg, Denmark. [doi: 10.1109/SmartGridComm.2018.8587464]
30. Le Cam L, Yang GL. Asymptotics in Statistics: Some Basic Concepts. New York, NY: Springer; 2000.
31. Gomatam S, Karr A, Sanil A. Data swapping as a decision problem. J Off Stat 2005;21(4):635-655 [FREE Full text]
32. Derpanis KG. The Bhattacharyya Measure. CiteSeerX. 2008. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.217.3369 [accessed 2021-11-01]
33. Joe H. Dependence Modeling with Copulas. New York: Chapman and Hall/CRC; 2015.

34. Borji A. Pros and Cons of GAN Evaluation Measures. arXiv. 2018. URL: http://arxiv.org/abs/1802.03446 [accessed 2020-05-22]
35. Kantorovich LV. Mathematical Methods of Organizing and Planning Production. Management Science 1960 Jul;6(4):366-422. [doi: 10.1287/mnsc.6.4.366]
36. Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks. 2017 Presented at: The 34th International Conference on Machine Learning; August 6-11, 2017; Sydney, Australia p. 214-223.
37. Ni H, Szpruch L, Wiese M, Liao S, Xiao B. Conditional Sig-Wasserstein GANs for Time Series Generation. SSRN. 2020. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3623086 [accessed 2021-11-01]
38. Friedman J. On Multivariate Goodness-of-Fit and Two-Sample Testing. 2003 Presented at: PHYSTAT2003; September 8-11, 2003; Stanford, California.
39. Hediger S, Michel L, Näf J. On the Use of Random Forest for Two-Sample Testing. arXiv. 2020. URL: http://arxiv.org/abs/1903.06287 [accessed 2020-05-06]
40. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55. [doi: 10.1093/biomet/70.1.41]
41. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. Circ Cardiovasc Qual Outcomes 2019 Jul;12(7):e005122 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.118.005122] [Medline: 31284738]
42. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. 2017 Presented at: Machine Learning for Healthcare Conference; August 18-19, 2017; Boston URL: http://proceedings.mlr.press/v68/choi17a/choi17a.pdf
43. Salim J. Synthetic Patient Generation: A Deep Learning Approach Using Variational Autoencoders. arXiv. 2018. URL: http://arxiv.org/abs/1808.06444 [accessed 2021-08-06]
44. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
45. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: Oxford University Press; 2004.
46. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. 2006 Presented at: 23rd International Conference on Machine Learning (ICML '06); June 25-29, 2006; Pittsburgh. [doi: 10.1145/1143844.1143874]
47. Hand J, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45(2):171-186. [doi: 10.1023/A:1010920819831]
48. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. 2019 Presented at: Advances in Neural Information Processing Systems 32 (NeurIPS 2019); December 8-14, 2019; Vancouver, BC p. 11 URL: https://papers.nips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html
49. Ping H, Stoyanovich J, Howe B. DataSynthesizer: Privacy-Preserving Synthetic Datasets. 2017 Presented at: The 29th International Conference on Scientific and Statistical Database Management; June 27-29, 2017; Chicago, IL p. 1-5 URL: https://dl.acm.org/doi/10.1145/3085504.3091117 [doi: 10.1145/3085504.3091117]
50. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput Stat Data Anal 2011 Dec;55(12):3232-3243. [doi: 10.1016/j.csda.2011.06.006]
51. Arslan RC, Schilling KM, Gerlach TM, Penke L. Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. J Pers Soc Psychol 2021 Aug;121(2):410-431. [doi: 10.1037/pspp0000208] [Medline: 30148371]
52. Bonnéry D, Feng Y, Henneberger AK, Johnson TL, Lachowicz M, Rose BA, et al. The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. J Res Educ Eff 2019 Aug 02;12(4):616-647. [doi: 10.1080/19345747.2019.1631421]
53. Sabay A, Harris L, Bejugama V, Jaceldo-Siegl K. Overcoming small data limitations in heart disease prediction by using surrogate data. SMU Data Science Review 2018;1(3):12 [FREE Full text]
54. Freiman M, Lauger A, Reiter J. Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau. United States Census Bureau. 2017. URL: https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/2017%20Data%20Synthesis%20and%20Perturbation%20for%20ACS.pdf [accessed 2021-11-01]
55. Nowok B. Utility of synthetic microdata generated using tree-based methods. 2015 Presented at: UNECE Statistical Data Confidentiality Work Session; October 5-7, 2015; Helsinki, Finland URL: https://unece.org/statistics/events/SDC2015 [doi: 10.1007/springerreference_64338]
56. Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. JPC 2018 Feb 02;7(3):67-97. [doi: 10.29012/jpc.v7i3.407]
57. Nowok B, Raab GM, Dibben C. Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R. SJI 2017 Aug 21;33(3):785-796. [doi: 10.3233/sji-150153]
58. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. eLife. 2020 Mar 11. URL: https://elifesciences.org/articles/53275 [accessed 2020-11-01]

59. Wang Z, Myles P, Tucker A. Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility Patient Privacy. 2019 Presented at: IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); June 5-7, 2019; Cordoba, Spain URL: https://ieeexplore.ieee.org/document/8787436 [doi: 10.1109/cbms.2019.00036]
60. Chin-Cheong K, Sutter T, Vogt JE. Generation of Heterogeneous Synthetic Electronic Health Records using GANs. 2019 Presented at: Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019; Vancouver, BC URL: https://www.research-collection.ethz.ch/handle/20.500.11850/392473 [doi: 10.3929/ethz-b-000392473]
61. Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and evaluation. J Am Med Inform Assoc 2020 Jan 01;27(1):99-108 [FREE Full text] [doi: 10.1093/jamia/ocz161] [Medline: 31592533]
62. Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences, 2nd ed. New York: Mcgraw-Hill Book Company; 1988.
63. Pihur V, Datta S, Datta S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. Bioinformatics 2007 Jul 01;23(13):1607-1615. [doi: 10.1093/bioinformatics/btm158] [Medline: 17483500]
64. Chen D, Yu N, Zhang Y, Fritz M. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. 2020 Presented at: ACM SIGSAC Conference on Computer and Communications Security; November 9-13, 2020; USA Virtual URL: https://dl.acm.org/doi/10.1145/3372297.3417238 [doi: 10.1145/3372297.3417238]
65. Hilprecht B, Härterich M, Bernau D. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. Proc Priv Enh Technol 2019;4:232-249. [doi: 10.2478/popets-2019-0067]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
GAN: Generative Adversarial Network
LR: logistic regression
pMSE: propensity mean squared error
SDG: synthetic data generation

Edited by C Lovis; submitted 15.12.21; peer-reviewed by H Turbe, N Zamstein; comments to author 04.01.22; revised version received 27.01.22; accepted 13.02.22; published 07.04.22.

Please cite as:
El Emam K, Mosquera L, Fang X, El-Hussuna A
Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study
JMIR Med Inform 2022;10(4):e35734
URL: https://medinform.jmir.org/2022/4/e35734
doi:10.2196/35734
PMID:35389366

Original Paper

Big Data Health Care Platform With Multisource Heterogeneous Data Integration and Massive High-Dimensional Data Governance for Large Hospitals: Design, Development, and Application

Miye Wang[1], MSc; Sheyu Li[2], PhD; Tao Zheng[1], BSc; Nan Li[1], MSc; Qingke Shi[1], MSc; Xuejun Zhuo[1], BSc; Renxin Ding[1], BSc; Yong Huang[1], MSc

[1]Engineering Research Center of Medical Information Technology, West China Hospital of Sichuan University, Ministry of Education, Chengdu, Sichuan Province, China

[2]Department of Endocrinology and Metabolism, MAGIC China Centre, Cochrane China Centre, West China Hospital, Sichuan University, Chengdu, China

Corresponding Author:
Yong Huang, MSc
Engineering Research Center of Medical Information Technology
West China Hospital of Sichuan University
Ministry of Education
Information Center, West China Hospital
No. 37, Guoxue Road, Wuhou District
Chengdu, Sichuan Province, 610041
China
Phone: 86 18980601030
Email: huangyong@wchscu.cn

Abstract

Background: With the advent of data-intensive science, a full integration of big data science and health care will bring a cross-field revolution to the medical community in China. The concept *big data* represents not only a technology but also a resource and a method. Big data are regarded as an important strategic resource both at the national level and at the medical institutional level, thus great importance has been attached to the construction of a big data platform for health care.

Objective: We aimed to develop and implement a big data platform for a large hospital, to overcome difficulties in integrating, calculating, storing, and governing multisource heterogeneous data in a standardized way, as well as to ensure health care data security.

Methods: The project to build a big data platform at West China Hospital of Sichuan University was launched in 2017. The West China Hospital of Sichuan University big data platform has extracted, integrated, and governed data from different departments and sections of the hospital since January 2008. A master–slave mode was implemented to realize the real-time integration of multisource heterogeneous massive data, and an environment that separates heterogeneous characteristic data storage and calculation processes was built. A business-based metadata model was improved for data quality control, and a standardized health care data governance system and scientific closed-loop data security ecology were established.

Results: After 3 years of design, development, and testing, the West China Hospital of Sichuan University big data platform was formally brought online in November 2020. It has formed a massive multidimensional data resource database, with more than 12.49 million patients, 75.67 million visits, and 8475 data variables. Along with hospital operations data, newly generated data are entered into the platform in real time. Since its launch, the platform has supported more than 20 major projects and provided data service, storage, and computing power support to many scientific teams, facilitating a shift in the data support model—from conventional manual extraction to self-service retrieval (which has reached 8561 retrievals per month).

Conclusions: The platform can combine operation systems data from all departments and sections in a hospital to form a massive high-dimensional high-quality health care database that allows electronic medical records to be used effectively and taps into the value of data to fully support clinical services, scientific research, and operations management. The West China Hospital of Sichuan University big data platform can successfully generate multisource heterogeneous data storage and computing power. By effectively governing massive multidimensional data gathered from multiple sources, the West China Hospital of Sichuan University big data platform provides highly available data assets and thus has a high application value in the health care field.

The West China Hospital of Sichuan University big data platform facilitates simpler and more efficient utilization of electronic medical record data for real-world research.

(JMIR Med Inform 2022;10(4):e36481) doi:10.2196/36481

KEYWORDS

big data platform in health care; multisource; heterogeneous; data integration; data governance; data application; data security; data quality control; big data; data science; medical informatics; health care

Introduction

Background

Emerging technologies, such as big data, the Internet of Things, cloud computing, and artificial intelligence, are profoundly changing medical and health care service models. Health and medical data sets, which are typically large with rapid growth, diverse data structure, multidimensional value density, high requirements of data credibility, and high concerns over data security, are often called health care big data. The term *big data* is used to represent not only a technology but also a resource and a method—a big data platform is a system and a tool that integrates data, tools, apps, and service.

Driven by national policies, high-income countries in Europe, such as the United Kingdom [1], and in North America, such as the United States [2], took the lead in building big data platforms for health care [3]. For example, Britain's comprehensive platform integrates and applies data from 12 categories, including health, medical care, transportation, and environment, to support government decision-making [3]. The Big Data Analytics Platform built by the Czech Republic meets the data analysis requirements of its national public health service [4]. In recent years, China has built a national-level health information platform that can be connected to provincial-level health information platforms for data integration [5].

In addition to national-level big data platforms in health care that are driven by policies, large health care institutions have also built big data platforms to suit their own management needs. At present, a medium-sized health care institution in China generates 1 to 20 TB of health care data, and a large health care institution generates 300 TB to 1 PB of health care data every year. Some hospitals have used big data technology to build hospital-level scientific research platforms—for example, Ninewells Hospital and Medical School in the United Kingdom developed a research data management platform [6], Asan Medical Center in South Korea developed a clinical trial management system based on integrated data [7], and the People's Hospital of Peking University in China developed a hospital-wide big data platform for clinical research [8]—or to build hospital-wide data integration platform—for example, the Third Hospital of Peking University [9] and the Second Affiliated Hospital of Guangzhou University of Traditional Chinese Medicine [10]. Most hospitals use big data technology in analytics platforms for diseases (eg, for nasopharyngeal carcinoma [11]), gastrointestinal conditions [12], cancer [13], and cardiomyopathies [14].

Meta-analyses [15-18] have found that big data platforms in health care have had great impacts on have had great impacts on medical technology, medical service quality, and medical costs, but the actual construction process is not easy, with challenges arising in data structurization, security, standardization, storage, processing, and management [15,16]. As a result, the value mined from this type of data is currently limited [17], and in China, health care data utilization is still in the early stage [18], which necessitates the improvement of health care big data governance.

Objectives

Health care data generated in large hospitals include (1) electronic medical records of clinical diagnosis and treatment, which contain diagnoses, prescriptions, surgical treatments, and examination findings; (2) data derived from health management or clinical research activities, including follow-up information, gene sequencing data, and physical examination data; (3) data related to hospital management, including patient wait times, bed turnover rate, medical equipment utilization efficiency, and revenue; and (4) data derived from web-based diagnosis and treatment services [19]. To make better use of medical data, the first task is to build a big data platform for data collection and governance to generate high-quality data assets. This will enable in-depth data analysis and mining, as well as the formation of rules of knowledge, allowing big data methods to benefit clinical practice, scientific research, and hospital management.

In addition to the technical considerations for platform construction, data management in large hospitals must take into account by data service management. Special attention should be paid to patient privacy protection and ethical concerns on data use. A hospital data platform is generally built for a specific purpose, such as scientific research, operations, data integration, or analysis, but large hospitals often require a comprehensive data platform—one that meets needs for medical treatment, education, scientific research, and management.

The overall objective of this study was to develop and implement a health care big data platform for a large hospital in China, with data governance as the core concept, to solve difficulties related to integration, calculation, storage, standardization, and security for multisource heterogeneous medical data. This platform will integrate the data from all operation systems in a hospital to generate high-quality data assets and form a massive high-dimensional medical database that can comprehensively support the clinical activities, scientific research, and management of the hospital.

From a technical aspect, construction of a big data platform needs to solve the following problems that are unique to the health care industry: (1) integration of multisource

heterogeneous data from multiple independent information systems within the hospital; (2) computing power requirements brought by the development of machine learning and deep learning during data application; (3) difficulties in analyzing and utilizing information data (due to nonuniform data standards, inconsistent use of a master patient index, and the fact that most electronic medical records in China are written in natural language, it is impossible to directly analyze and use existing information data, and as a result, semantic interoperability must be improved with the use of medical terminology) and (4) data security and patient privacy protection.

Methods

Overview

The West China Hospital of Sichuan University (WCH) built a health care big data platform, which is referred to as the WCH-BDP or *the platform* hereafter. WCH is a world-renowned large hospital with more than 4800 beds and approximately 15,000 outpatients on average per day. WCH's electronic medical record system was built in 2007; therefore, it has been in use for 14 years. The hospital has more than 100 departments and sections, and their clinical activities generate a massive amount of data.

Traditional information technology can no longer handle massive amounts of data that are continuously growing, causing difficulties in effectively integrating multisource heterogeneous data, the serious problem of isolated data islands, bottlenecks in data storage and calculation, difficulties in structurally utilizing medical records written in Chinese semantics, and high technical barriers in mining data from images, videos, and files. In 2017, WCH launched a hospital-wide health care big data platform project to address these difficulties.

The project focuses on the design and development of the platform architecture and does not involve any study of clinical data, so ethical declarations are not applicable.

Project Organization

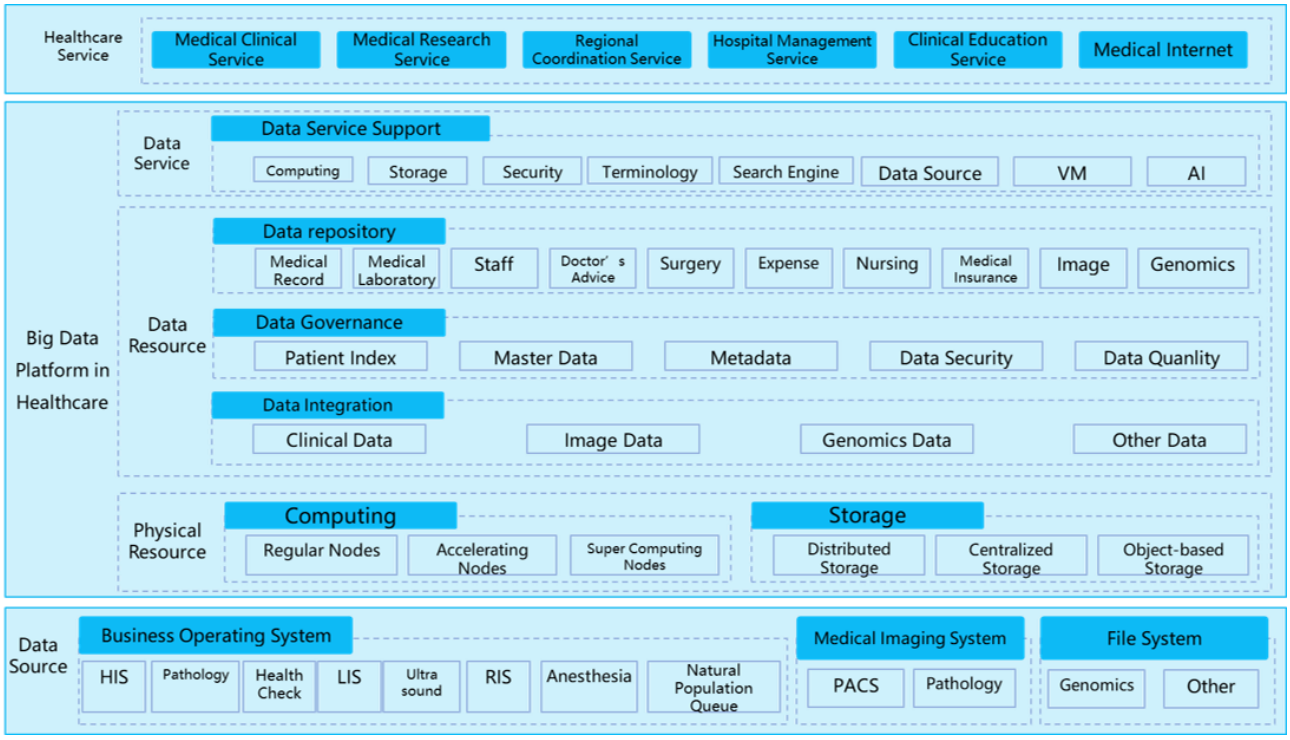
The first step of platform construction was organizing a project management team. To ensure the performance of the platform, WCH set up 2 working groups—one for platform construction, and the other for platform management. The construction working group included the chief platform architect, information technology experts, system engineers, and data engineers. The hospital information center was responsible for building the platform. The construction working group focused on the following: (1) the application objectives of the platform, (2) data integration method and scope, (3) master patient index strategy, (4) reference standards for medical terms, (5) scope of the master data, (6) data model structure, and (7) system implementation and training. Project meetings were held weekly and status meetings were held monthly to summarize the progress of the phases until the platform was launched. The discussions in these meetings provided a solid foundation for platform design and development. After the launch of the platform, the construction working group also conducted regular training on system functions, maintenance, and the help manual.

The management working group included all stakeholders, such as the chief information officer, hospital management users, clinical users, and scientific research users. The management group was the data governance committee responsible for organizing and overseeing all data-related work, including data definitions, data benchmarking, data quality control, and data security, after the launch of the platform. The committee was in charge of formulating related management systems, work collaboration mechanisms, and procedure standardization.

WCH-BDP Framework Design Strategy

The convergence of multisource heterogeneous data from all operation systems in the hospital was necessary to ultimately provide data to different medical services (Figure 1). Therefore, the design strategy was to use computing and storage devices with enough capacity to integrate data into the corresponding physical resources, with separate storage and computing processes based on the characteristics of the modality (eg, clinical data, image data, or genomic data). A data repository was built using data governance methods to meet the needs of different subject fields. The platform had to be able to combine all data service supports, such as data security service, terminology service for data governance, search engine service, virtualization service, and artificial intelligence service—all of which rely on the computing power and storage capacity of physical devices and data resources. Solutions for data integration and governance were core aspects of the construction process. Data governance, which is the core of data management and the basis for standardizing disordered data into highly available data, includes master index data governance, master data governance, metadata governance, data security management, and data quality control management [10,20,21].

Figure 1. Big data platform architecture.



Data Storage and Calculation

Massive data were collectively stored in the storage devices of the WCH-BDP. To better support data analysis, storage and calculation environments were designed and provided according to the characteristics of each data modality. Generally, the amount of data structuration is negatively correlated with storage space and computing power (ie, the greater the data structuration, the lower the requirements for storage space and computing power) (Figure 2). Hence, data that are poorly structured require more storage space and stronger computing power.

Highly structured data do not require much storage space. They can be stored on a distributed storage device, and computations can be effectively run by conventional central processing units.

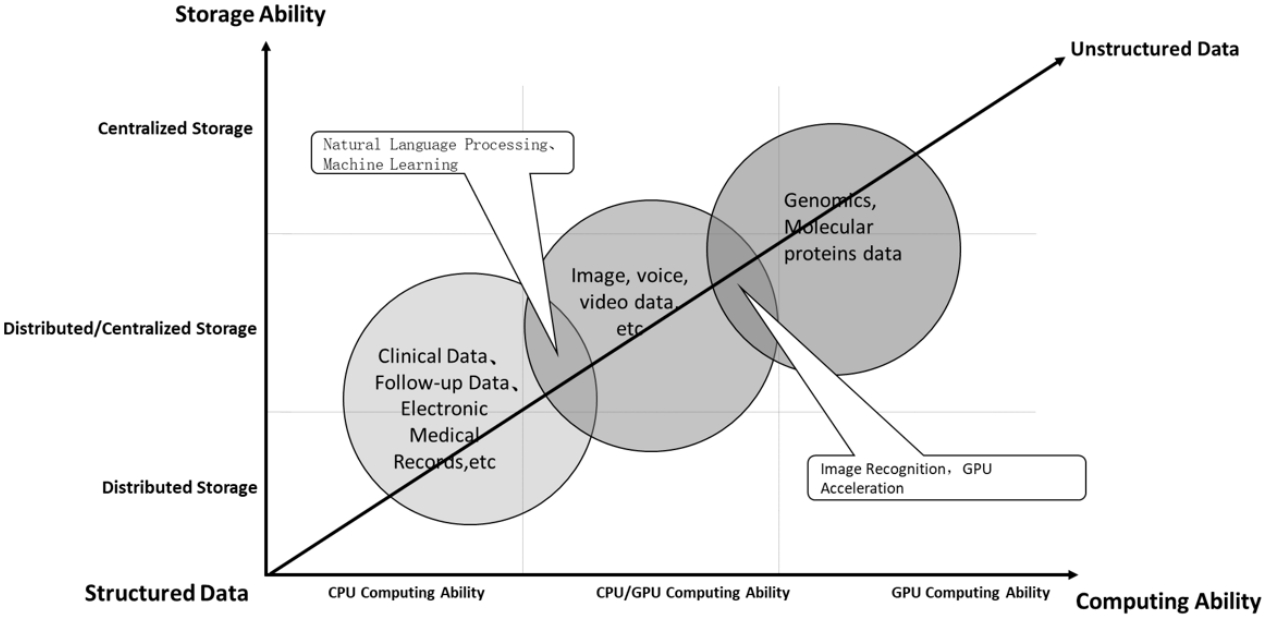
Semistructured data, such as the records of a patient's major complaint, disease history, examination findings, and examination conclusions in Chinese electronic medical records do not occupy much storage space and can be stored in distributed storage devices. However, these data require natural language processing for analysis; therefore, both general graphics processing units (GPUs) and central processing units are needed to provide sufficient computing power.

Unstructured data that are large in total volume but small in individual volume (eg, the original image data in DICOM format generated by thin-layer scanning examinations such as radiology and ultrasound) require a relatively high storage space and must be stored in centralized storage devices (mostly network-attached storage) to save resources. To analyze and mine these data, characteristic modeling is performed mainly with machine learning or deep learning technologies, which requires the support of a large amount of GPU power and private GPU resources.

Unstructured data with a large individual volume, such as genomic data, are generally stored directly in object-based storage. Genomic sequencing generates not only a large amount of original sequencing data, but also, a large amount of data related to multiple processes in research on biological information. As a result, genomic data sets often have a tremendous size, and the storage space that is required is mostly measured in petabytes. Analysis and mining of the types of data stored in object-based storage require a lot of clustered computing power supplied by multiple high-performance GPUs, which necessitates the deployment of accelerated supercomputing power supported by GPU clusters.

Traditional big data platforms are planned and implemented in a unified way with storage and computing integration. The design of the WCH-BDP is the realization of storage and computation separation. The WCH-BDP first solves the problem of effective storage and management of massive genomic data files. Second, when researchers use genetic data for analysis, the scheduling software provided by the WCH-BDP loads the analysis data into the supercomputer environment for analysis. After the analysis is completed, the analysis results are retained, and the temporary storage space occupied by the analysis process is released.

Figure 2. The relationship between the degree of data structuration and storage and computing capability.



Data Integration

In China, medical institution operations information database systems mainly include Microsoft SQL Server, Oracle database, MySQL database, Caché database, and MongoDB; other unstructured data are mostly stored in the form of files. The WCH-BDP achieves data integration through both real-time and non–real-time data entries. For real-time data integration, a master–slave database is generated, and the log data of the slave database are parsed and captured in real time. Real-time data integration is suitable for an operations information system with mostly structured data. Parsing the information in the slave database does not consume much physical resources and therefore does not affect the performance of the master database, which ensures the security and stability of data integration. Image data integration and entry can be completed by directly reading image files using the DICOM protocol. Genomic data and other data in the form of files are not integrated in real time and are entered via file transfer protocol.

Master Index Governance

The enterprise master patient index is the unique tag of a patient in a health care institution. One patient might have different enterprise master patient indexes in the different operating systems of a hospital because these systems were constructed independently. This necessitated a master patient index governance to standardize the enterprise master patient indexes. The governance of the master index data is achieved through the master index governance system. The governance system included 4 stages: data preparation, standardization strategy, data processing, and tracking or feedback (Figure 3).

The data governance strategy of the WCH-BDP mainly focused on 3 key values: ID number, name, and telephone number. The configurations and data processing of these 3 key values are shown in the following table (Table 1).

Master index data processing cannot rely only on the system for automatic completion. For example, in the fifth row of Table 1, both the Name and Telephone number are *Equal*, but the ID card number is *Unequal*, which might be because the same patient used a different ID card for registration. In this case, the last step for standardization of the master index is critical, during which the uniqueness of each master index can be ensured by manual adjustment after log analysis or by referring to the actual operational process.

Figure 3. Flowchart of master index governance.

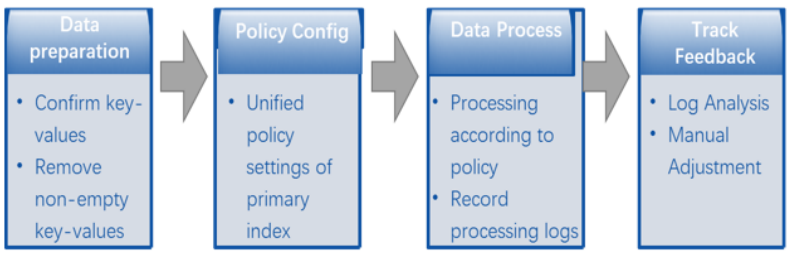


Table 1. Governance strategy of enterprise master patient index.

ID number	Name	Telephone number	Result
Equal	Equal	Equal	Accept
Equal	Equal	Unequal	Accept
Equal	Unequal	Equal	Accept
Equal	Unequal	Unequal	Denied
Unequal	Equal	Equal	Accept
Unequal	Equal	Unequal	Denied
Unequal	Unequal	Equal	Denied
Unequal	Unequal	Unequal	Denied

Master Data Governance

Master data include all related data items listed in the data dictionary, such as health care institution code, drug code, diagnosis code, and anesthesia method [22] (Table 2). Master data governance is the aim to map and handle master data discrepancies caused by different system standards. Classification and reference standards for the master data to be processed are determined, and they are used to map relationships between data in the database. The results are published to users for subscription and application.

Table 2. Example of the WCH-BDP master data reference standard.

Classification of master data	Number of reference standards	Example
Classification of diseases	5	ICD-10GB/*t* 14396-2016 Classification and codes of diseasesGB/*t* 15657-1995 Classification and codes of diseases and ZHENG of traditional Chinese medicine
Basic industry information	6	GB 11714-1997 Rules of coding for the representation of organizationGB/*t* 13745-2009 Classification and code of disciplinesGB/*t* 2260-2007 Codes for the administrative divisions of the People's Republic of China
Health informatics	20	GB/*t* 21715-2020 Health informatics—Patient healthcard dataGB/*t* 24465-2009 Health informatics. Health indicators conceptual frameworkGB/*t* 25512-2010 Health informatics-guidelines on data protection to facilitate trans-border flows of personal health informationGB/*t* 30107-2013 Health informatics.HL7 Version 3.Reference information modelGB/Z 24464-2009 Health informatics-Electronic health record-definition, scope and contextGB/Z 28623-2012 Health informatics. Interoperability and compatibility in messaging and communication standards. Key characteristics
Personal information	12	GB/*t* 2261-2003 Classification and codes of basic personal informationGB/*t* 4658-2006 Codes for record of formal schoolingGB/*t* 4761-2008 Codes for family relationshipGB/*t* 6565-2009 Classification and codes of occupationsGB/*t* 8561-2001 Code of professional technical position
Information technology	3	GB/*t* 34960.1-2017 Information technology service－Governance GB/*t* 39725-2020 Information security technology—Guide for health data security

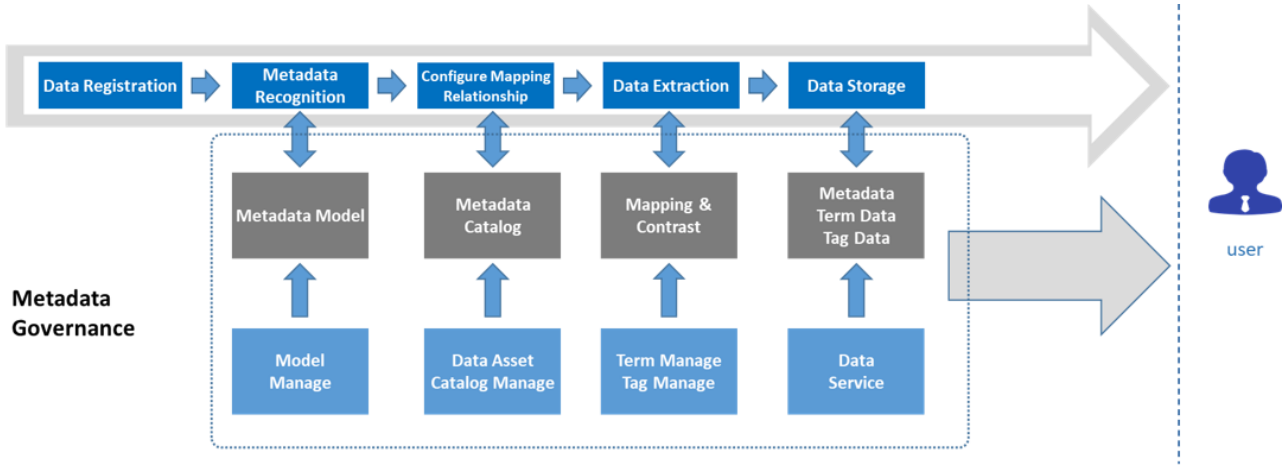
Metadata Governance

Metadata Governance Process

Metadata are the bridge between the data and the data users. They describe the content (what), coverage (where, when), quality, management method, owner (who), and provision method (how) of data. The metadata governance approach used by the WCH-BDP is shown in Figure 4.

In the WCH-BDP, a unified metadata model for identifying the registered original data was designed and constructed based on the standards of the Common Warehouse Metamodel of the Object Management Group. Based on the metadata directory, the relationship for data mapping is configured to initiate data extraction. The mapping results are saved in terms of the configured relationship and compared with the standard terminology and tags in the metadata management system. All extracted data are stored in the platform's data repository, which provides data services to the application layer using standardized metadata, terminology, and tags.

Figure 4. Metadata governance architecture.



Terminology Standardization

Terminology standardization is an important step in metadata governance. The WCH-BDP uses medical terminology from the Chinese Open Medical and Healthcare Alliance to standardize the terminology in the metadata system.

Governance of Tag Data

Tagging is also one of the more effective methods used for data governance. Unlike terminology synonyms, the *tag* is an attribute that reflects the application of data [23]. Tag data are more suitable for the preprocessing scenario of comparing or calculating a large amount of data, such as the change in the value of a test result before versus after surgery, or the range of patient blood pressure change before and after taking a medication. The governance of tag data should consider the app objectives and continuously track and maintain tag data.

Natural Language Processing Governance

Many medical records and reports with descriptions written in Chinese are difficult to process due to the unique characteristics of the Chinese language. Currently, in China, one mainstream post–structured data processing method is the use of machine learning technology combined with health care terminology to segment words and refine concepts from Chinese text [11]. It is common practice to optimize the word segmentation algorithm model by comparing the word segmentation results with standard terms and aggregating the comparison results, through which a comprehensive corpus and a more accurate word segmentation algorithm can be developed. The algorithm model and the content of the corpus may depend on disease type, regional culture, and even writing habits; therefore, these factors had to be considered during construction.

Data Security Management

Overview

Data utilization and security are contradictory. To ensure proper and secure data use, all countries have issued laws and regulations on the security of health and medical data. Referring to relevant international and domestic laws and regulations on security protection, the WCH-BDP uses a *5S* (ie, data security, app security, use security, management security, and ownership security) control system for data storage, access, use, and transformation of scientific and technological achievements.

Data Security

Data security refers to data classification and hierarchical management security. Scholars from Harvard University have proposed that a data classification system should be established to ensure data security according to the laws and regulations of Health Insurance Portability and Accountability Act and related ethical regulations on scientific research [24,25]. In the WCH-BDP, data security has 5 levels, with 3 dimensions: affected object, degree of impact, and scope of impact. The data access interface with corresponding restrictions is provided to each level accordingly (Table 3).

Table 3. Data security strategy of the big data platform.

Security classification	Security level	Descriptions	Examples	Precautions
Level 1	Most confidential	Hypersensitive information	Financial data, personal authentication data	A specified environment for a specified single individual to use
Level 2	Confidential	Highly sensitive information	Credit data, personal health privacy data	A specified environment for a specified single role to use
Level 3	Secret	General sensitive information	Personal information during diagnosis and treatment, contract information, employee management data of employees	Use after the role group is authorized
Level 4	Internal use only	Information that is not publicly disclosed	Organizational structure, basic information of employees, general data after desensitization	Use after internal authorization
Level 5	Open to the public	Data that can be publicly disclosed	Summarized results obtained by statistical analysis	Public access or use

App Security

App security refers to the data app security management of a system during its processes. In the WCH-BDP, a unified master data authentication service is provided for user authentication so that each app system manages only role and app permissions required by the system itself. When an actual role user uses an app, data desensitization and encryption are completed through the unified platform service interface to ensure the security of the data app.

Use Security

Use security refers to the security management during the processes of data processing, use, and analysis. The National Institutes of Health has emphasized that special attention should be paid to privacy protection in data apps [26]. Patient information that needs to be desensitized generally includes all of the information that can be directly and indirectly linked to or used to locate the patient, including name, ID, telephone number, address, contact information, information related to infectious disease [27]. The WCH-BDP ensures patient privacy protection through desensitization, encryption (Table 4), and multiparty computation.

Multiparty computation is a controllable and measurable method proposed by a Turing Award winner, Qizhi Yao, to solve the problem of data abuse [28]. Before statistical analysis of massive data, the multiparty computation service gateways in the WCH-BDP are deployed to complete all data calculations in the security domain and provide the final calculation results to users through the platform service. This process can effectively manage data use security.

Table 4. Data desensitization and encryption policies of big data platform.

Name of strategy	Scope of data involved	Design of desensitization and encryption policy
Digital data	Operating revenue, key quantity...	Fuzzy rounding method or fuzzy percentage method
Structured data of fixed length	Identification card number, telephone number, name...	Replace or encrypt from the starting bit to the end of the specified length range
Text data of varying length	Address, electronic medical record, descriptive diagnosis (infectious disease)	Locating sensitive content, then replace or encrypt sensitive characters
Image data	Radiological, ultrasonic and pathological imaging data	In image files, encryption algorithm is used for desensitization and watermarking is configured
File data	Genomics, molecular protein data	Locating sensitive content and rename sensitive characters

Management Security

Management security of platform data requires that all platform services, including data services, be completed under secure closed-loop management. In the WCH-BDP, all tasks—data access, management, and governance; data analysis, utilization, and mining; and the transformation and realization of scientific research findings—are completed through platform services for data resources, storage, computing power, and network access. Establishment of the management security system can effectively guarantee the security of the data utilization process.

Ownership Security

Currently, most research findings are obtained in a laboratory or scientific research setting but are difficult to apply in an actual production setting, which directly leads to a low transformation rate of research achievements and insufficient recognition of intellectual property rights. In the WCH-BDP, through the integrated service gateway that connects to the real service operation scenario, published research findings can be directly introduced into clinical practice, realizing the translation of scientific research to the clinic. In this process, the design of the integrated service gateway management system can effectively ensure researcher's ownership security.

Data Quality Management

Overview

Data quality management is a continuous process and an important link in data governance. Health care data quality management in the WCH-BDP involves the following steps: (1) building a data quality standard system; (2) evaluating current data quality; (3) analyzing data quality problems; (4) developing and optimizing the solutions in detail to solve problems; and (5) establishing a data quality control knowledge database for future reference. The steps above are continuously iterated to form a closed-loop system for health care data quality management.

Standards of Data Quality

Common data quality problems include data outliers, duplications, data without any clear relationships, orphaned data records, and data that make no logical sense medically. Fully referring to the evaluation and grading standards of electronic medical records issued by the National Health Commission of China and research findings by experts in the field of health care [29-31], we developed a standard system for data quality control using the dimensions *consistency*, *integrity*, *integrability*, *timeliness*, and *stability* as parameters for evaluation (Table 5).

Table 5. Data quality standard system.

Dimension	Content of the dimension	Quality indicator	Rule
Consistency	Check whether the data value is in the dictionary domain	Consistency rate	Higher than 90%
Integrality	Check the completeness of the required data	Percentage of integrality	Higher than 80%
Relevance	Check the relationship between key data	Degree of relevance	Higher than 95%
Timeliness	Check the logical validity of time-type data	Timeliness ratio	Higher than 80%
Uniqueness	Check whether data duplication exists	Repetitive rate	Lower than 0.01%
Stability	Check whether the fluctuation of data volume is abnormal	Fluctuation ratio	Lower than 20%

Process Quality Control

In addition to data quality control for original data, quality control management of procedures in operations are included [21]. Data quality problems reflect problems in operations processes. By formulating a plan that optimized data quality by improving management and operational protocols, the WCH-BDP achieved its goal of controlling the quality of operations to obtain high-quality data outputs. Data quality management is a continuous optimization process that requires the participation of all personnel involved.

Data Service Support

The WCH-BDP is an environment that can effectively support data services with voluminous data repositories and strong physical resources for data computation and storage. Under an overall ecology of data security management, the platform can provide support to data services such as computing, storage, and virtualization. Search engine services, terminology services, and artificial intelligence services can be effectively integrated to meet the needs of all health care operations.

Results

General

The WCH-BDP was formally launched in November 2020, after 3 years of construction, for demand surveys, platform design, module development, and pilot operation. To date, the platform has formed a massive high-dimensional database with more than 12.49 million patients, 75.67 million visits, and 8475 data variables. The platform has brought hospital informatization to a new level and greatly improves the hospital's overall big data capabilities.

Computing Power

The WCH-BDP has more than 20 PB data storage capacity and runs on a server cluster consisting of 252 physical servers: 117 servers are dedicated to computing and analyzing, and 38 are equipped with more than one GPU card. In the server cluster dedicated to computing, there are 320 CPU cores and 149 GPU cards. In total, the platform has computing powers that exceed 300 TFLOPS for clinic data, 600 TFLOPS for image data, and 1000 TFLOPS for genomic data.

Scope of Data Integration

The WCH-BDP integrates diagnosis and treatment data from clinical information systems, clinical research data from science and research information systems, and management activity data from management information systems. Of 134 hospital systems, the platform has integrated the data of 27 systems, including hospital information systems, laboratory information systems, radiology, ultrasound, web-based medical appointments, human resources, equipment and supply management, Internet medical service since January 2008. In addition, all image data in the picture and communication systems and more than 20,000 patients' genomic data have been entered into the platform.

Database

Using Fast Healthcare Interoperability Resources Reference Information Model 3.0 standards, data integrated into the WCH-BDP were reorganized and arranged based on the type of operational activities while considering the characteristics of China's medical system. Finally, a database that includes 134 data charts for 18 subject fields was developed. The top 5 subject fields that were responsible for the most data in the database are listed in Table 6. To date, the database includes 8475 data variables and 6.272 billion rows of data records.

Table 6. Top-5 subject fields in terms of number of data rows.

Number	Subject field	Tables, n	Data variables, n	Rows (10,000×n)	Systems involved
1	Medical record	10	476	369824.53	Hospital information system, online diagnosis and treatment system
2	Medical technology	6	502	115073.35	Electrocardiogram system, radiology information system, echocardiography system, endoscopic system, dynamic electrocardiogram system, pathology information system, ultrasonic system, laboratory information system, interventional surgical workstation, medical technology reservation information system, outpatient information system, physical examination information system
3	Fee	8	440	60855.24	Hospital information system, physical examination information system, online diagnosis and treatment system
4	Medical advice	3	395	21366.64	Hospital information system, online diagnosis and treatment system
5	Staff	13	802	18138.94	Hospital information system, physical examination information system, electronic data capture, human resource system

Data Asset Directory

The platform has constructed a data asset directory for users by further standardizing the data through the governance of master data, metadata, terminology, and natural language processing. To ease the process of inquiry, the directory has a hierarchical tree structure that allows users to select the data items that they need using fuzzy searches. To minimize comprehension difficulties, the directory uses a data variable–naming scheme consistent with the names in the interface of the corresponding operations systems as much as possible; in addition, users can see the information of the data source, content, and range of values.

The data asset directory covers 13 fields and 1488 data variables (Table 7). Of the 13 fields, the original data variables in 9 fields are structured variables, and those in 3 fields (medical records, imaging examinations, and nursing care records) are semistructured text. Semistructured data were converted into poststructured derived variables by using approaches such as word segmentation, entity extraction, and semantic identification. For patient tag information, the app-oriented poststructured derived variables were generated using a data-mining algorithm.

Table 7. List of data assets.

Number	Directory field	Data variable (unit)	Types of variables	Example
1	Demographic	91	Structured	Gender, age, occupation, present address, nationality, height, weight, blood type
2	Basic medical information	410	Structured	Appointment date, visit date, clinic department, clinic type, supervising doctor, the department transferred to, admission date, discharge date, discharge state
3	Medical record information	123	Unstructured	Admit note, progress note, discharge record
4	Clinical diagnostic information	47	Structured	Clinic diagnosis, emergency diagnosis, admitting diagnosis, discharge diagnosis, medical insurance diagnosis, pathological diagnosis
5	Surgical and operational information	138	Structured	Name of operation, name of procedure, surgeon, surgical grade, anesthesia grading, incision type, level of healing
6	Diagnosis and treatment information	166	Structured	Type of medical order, drug name, usage, dosage, frequency, execution time of medical order
7	Laboratory test data	147	Structured	White blood cell count, red blood cell count, sodium level, uric acid level, blood glucose level, creatinine level
8	Imaging results	124	Unstructured	Magnetic resonance imaging, computed tomography, x-ray, ultrasound, digital radiography
9	Nursing record information	53	Unstructured	Admission assessment, daily records, nursing records
10	Physiological monitoring data	50	Structured	Vital signs
11	Scale evaluation data	50	Structured	Mood index, pressure ulcer assessment, risk assessment of falling out of bed
12	Medical cost information	65	Structured	Name of charge items, amount of charge items, settlement time
13	Patient label information	24	Unstructured	Patients with lower test results after surgery, patients with higher blood pressure after medication

Assessment of WCH-BDP Performance

Service Support to Major Projects

Supported by the data services of the platform, the research teams of WCH have constructed more than 120 disease databases for different research aims, including a number of national-level multicenter disease databases. Since it was launched, the platform has supported more than 20 clinical and hospital management research projects, of which, 3 have won second prize in the State Science and Technology Progress Award of China and first and second prizes in the Sichuan Provincial Science and Technology Progress Award.

The platform's support of basic research is exemplified by a project on the mechanisms of allosteric regulation and signal transduction of G protein–coupled receptors conducted by the State Key Laboratory of Sichuan University; using the storage and computing power provided by the WCH-BDP, the research team revealed the microconversions of key amino acids during allosteric regulation, which laid the foundation for the design and screening of G protein–coupled receptor–targeting small molecule allosteric regulators (unpublished, X. Yang, PhD, 2022).

Another example of the platform's support of clinical research is a lung cancer research project at WCH. Using the data resources, storage resources, computing power, and exploration environment provided by the platform, the lung cancer research team identified and validated high-sensitivity high-specificity markers for early diagnosis of lung cancer [32], and the team further developed the first lung cancer database and artificial intelligence–assisted product for lung nodule diagnosis in China [33]; these papers [32,33] are listed in Essential Science Indicators—published in *Cell* (impact factor 41.582) and *Signal Transduction and Targeted Therapy* (impact factor 18.187). Research results have also been published in internationally renowned academic journals, such as *Medical Image Analysis* (impact factor 11.148) [34] and *Nature Biomedical Engineering* (impact factor 25.671) [35].

Artificial intelligence–assisted products for lung nodule diagnosis can detect 3 mm to 5 mm pulmonary nodules with 98.8% accuracy through artificial intelligence technology, which was significantly better than the performance of domestic and foreign specialists (79.9% of doctors with senior experience in Peking Union Medical College, only 40.9% of those with junior experience), and the average reading of each chest computed tomography (CT) can save 3 to 5 minutes [33]. In 2020, the system was used in more than 100 hospitals nationwide, including in West China Hospital of Sichuan University. It not only improves the efficiency of chest CT image reading, but also reduces the rate of missed diagnosis of small pulmonary nodules. It also plays an important role in realizing the homogeneity of early diagnosis of lung cancer.

Using this platform, research on segmentation of adrenal glands from CT images [36], in which a novel 2-stage deep neural

network for adrenal gland segmentation in an end-to-end fashion was proposed, used data resources, storage resources, computing capabilities, and the exploration environment provided by the WCH-BDP platform. The research data set contained 348 CT volumes acquired from 348 patients, which was used to verify the performance of the new method and show that the new cascaded framework outperformed, with respect to accuracy, state-of-the-art deep learning in segmenting the adrenal gland [36].

Changes in the Conventional Data Service Model

The launch of the WCH-BDP has led to tremendous changes in the data service model of the hospital (Table 8). Users used to rely on information systems personnel for data use, but now they can analyze data by themselves throughout the entire process. Using the search engine service provided by the platform, researchers can quickly retrieve data from the databases to form a disease database suitable for real-world research, which is then introduced into the information exploration environment for data statistical analysis and mining.

Table 8. Changes between the traditional and the present data service.

	Traditional data services	Data services base on the platform
Data visualization	Data not visible	Users can visually view the available data catalog
Data retrieval	Data engineers develop code through experience	Users can customize the search format and output format through the search engine and preview the results
Data approval	Data are available after ethical review and clinical study program approval	Data are available after ethical review and clinical study program approval
Data mining	Use your own computer to analyze data	Development environment and tools, such as R, SPSS, and Python, can be used on the platform, and computing power provided by the platform can be called by data mining algorithm
Data download and access	Download the data which data engineers develop and perform encryption	The platform creates accounts with different permissions for registered users. In a network environment after security authentication, authorized users can log in to the big data platform unified portal through virtual desktop infrastructure. The platform provides each authorized user with private storage space of different capacities. Users can directly store their research results in this space, or install the software developed by our college on their personal computers to transfer the research results to personal computers

The launch of the WCH-BDP has substantially improved the capacity of data services (Table 9). The scope of usable data in the platform is 3.37 times that of the previous level, an increase from 8 operation systems covered by the previous data warehouse to 27 systems by the big data platform. The dimensions of usable data are 1.8 times those of the previous level, an increase from 803 data variables to 1488 variables. The amount of usable data is 2.4 times the previous level, increasing from 6.8 billion rows of data records to 16.49 billion rows.

The engineers completed 996 instances of manual data service in the 6 months before the launch of the platform (monthly mean 166). In contrast, 8561 self-service data retrievals were completed each month after the launch of the platform—a 51-fold increase in service efficiency. Each person could complete 2 instances of data services each day before but could complete 65 instances per day after the launch of the platform due to the help offered by the automated search engine—a 37-fold increase. In addition, the platform shortened the average duration of each data service 30-fold, from 4.5 hours to 0.15 hours. The platform substantially improved the volume and efficiency of data services.

Table 9. Comparison of data service capabilities.

	Before the launch	After the launch
Number of business systems covered	8	27
Data dimension	803	1488
Data volume (billion)	6.8	16.49
Number of monthly services	166	8561
Time per request (hour)	4.5	0.15

Improvement of Data App Security

Data security is a key focus of the platform. In the past, data security relied mostly on the professional ethics of data engineers. In contrast, in this platform, which automatically manages data and provides data services, data security management is mostly system-based and manually assisted because all operation activities leave footprints in the system. This can prevent the abuse of individual permissions and effectively ensure data security.

Discussion

The performance of the WCH-BDP has demonstrated that data resources can be effectively converged and governed to form highly usable data assets that have extremely high application value in the field.

The success of the construction of a big data platform in health care is based on the following: (1) The project is managed with a strong organizational structure that has a top-down data governance committee involving multiple parties. The committee leads and oversees data governance duties. (2) The project is led by an information technology department that can provide technical support. The information technology team should have excellent skills and be familiar with the operations and data connotations of all systems in the hospital. (3) All departments and sections of the hospital should participate in the project, and detailed demand surveys and analyses should be performed. (4) The project requires sufficient scientific knowledge in medical informatics, management, and engineering to ensure smooth integration of medicine, management, and informatics for overall framework design. (5) The ethics office and clinical research management sections of the hospital should participate in the project to ensure patient privacy protection and data security. (6) The construction project needs suppliers who have rich experience and can provide sufficient technical support. (7) A sufficient amount of servers are needed for data storage and computation. (8) The project needs sufficient financial support.

Similarities between the WCH-BDP and other data platforms [1-14] are that they provide data services through data integration, need to complete data integration, have various medical data, can provide structured data services, and can provide massive data retrieval. However, there are 6 differences: (1) The WCH-BDP integrates all business system data, while most other platforms integrate data on demand. (2) Our platform parses database logs and migrates full business data to the data center in a master–slave database synchronization mode. Some of the platforms use API interfaces to implement data migration. (3) The WCH-BDP accesses data constantly and in real time. Nevertheless, some other data platforms often access data by the day. (4) While most other data platforms only integrate clinical data, the WCH-BDP integrates both clinical data and hospital management data. (5) Most other data platform may not have supercomputing capability. After integrating supercomputing capability, the WCH-BDP can store more than 20 PB and calculate faster than 1900 TFLOPS. (6) Most other platforms can only provide conventional data storage and processing functions by disease type. The WCH-BDP provides an analysis environment equipped with data mining tools, including open-source tools, such as R and Python, and paid apps, such as SAS and SPSS. Researchers can use distributed clusters for data mining.

The WCH-BDP can be further improved and optimized by (1) connecting more operations systems of the hospital to the platform and continuously optimizing the data governance strategies; (2) further utilizing and mining the data (eg, exploring multimodal artificial intelligence apps in health care); (3) making the platform a multicenter public platform by launching transdisciplinary, cross-hospital, cross-regional collaborations and including more medical information data; and (4) providing the hospital with further full-cycle *standardization* + *security* + *service* big data services.

Acknowledgments

This study was sponsored by the National Health Commission (Big data integration and application platform construction of West China Hospital of Sichuan University). Thanks are due to Engineering Research Center of Medical Information Technology, West China of Public Medical Information Services Co Ltd, Shanghai ClinBrain Co Ltd, and New H3C Technology Co Ltd for assistance with the development of data integration and data governance.

Conflicts of Interest

None declared.

References

1. Find open data. UK Government Digital Service. URL: https://www.data.gov.uk [accessed 2010-10-20]
2. U.S. Department of Health & Human Services. URL: http://www.healthdata.gov/ [accessed 2010-09-01]
3. Wu M, Zhen TM, Gu JL, He YQ, Mu Y, Song KM, et al. Development of health care big data at home and aboard and the application prospect in health decision support. Soft Sci Health 2019 Feb 17;33(2):76-79. [doi: 10.3969/j. issn. 1003－2800. 2019. 02. 017]
4. Štufi M, Bačić B, Stoimenov L. Big data analytics and processing platform in Czech Republic healthcare. Appl Sci (Basel) 2020 Mar 02;10(5):1705-1705.23. [doi: 10.3390/app10051705]
5. Zhou GH, Xu XD, Zhang XG, Hu JP. Design of data governance system based on national health information platform construction. Chin J Health Inform Manag 2019;16(2):131-134. [doi: 10.3969/j.issn.1672-5166.2019.02.02]
6. Nind T, Galloway J, McAllister G, Scobbie D, Bonney W, Hall C, et al. The research data management platform (RDMP): a novel, process driven, open-source tool for the management of longitudinal cohorts of clinical data. GigaScience 2018 Jul 01;7:1-12 [FREE Full text] [doi: 10.1093/gigascience/giy060] [Medline: 29790950]

7. Park YR, Yoon YJ, Koo H, Yoo S, Choi C, Beck S, et al. Utilization of a clinical trial management system for the whole clinical trial process as an integrated database: system development. J Med Internet Res 2018 Apr 24;20(4):e103-e103 [FREE Full text] [doi: 10.2196/jmir.9312] [Medline: 29691212]
8. Wu Y, Li MX, Ding YJ, Dong S, Liang GW, Wang BT. Development of a big medical database system for clinical research and practice of medical data governance. Chin J Med Sci Manage 2021 Apr 21;02(34):81-86. [doi: 10.3760/cma.j.cnl13565-20201012-00320]
9. Ji H, Li W, Jia M. Integration platform and data integration application based on big data. Chin J Health Inform Manag 2017 Aug 20;04(14):525-529. [doi: 10.3969/j.issn.1672-5166.2017.04.01]
10. Fu H, Xu F, Fan MY. Discussion on big data governance and system construction of hospital health care. Chin J Libr Inf Sci Tradit Chin Med 2019 Jun 15;03(43):1-5. [doi: 10.3969/j.issn.2095-5707.2019.03.001]
11. Lin L, Liang W, Li CF, Huang XD, Lv JW, Peng H, et al. Development and implementation of a dynamically updated big data intelligence platform from electronic health records for nasopharyngeal carcinoma research. Br J Radiol 2019 Oct 04;92(1102):92:20190255 [FREE Full text] [doi: 10.1259/bjr.20190255] [Medline: 31430186]
12. Yan L, Huang W, Wang L, Feng S, Peng Y, Peng J. Data-enabled digestive medicine: a new big data analytics platform. IEEE/ACM Trans Comput Biol and Bioinf 2021 May 1;18(3):922-931. [doi: 10.1109/tcbb.2019.2951555]
13. Cha HS, Jung JM, Shin SY, Jang YM, Park P, Lee JW, et al. The Korea cancer big data platform (K-CBP) for cancer research. Int J Environ Res Public Health 2019 Jun 28;16(13):16, 2290 [FREE Full text] [doi: 10.3390/ijerph16132290] [Medline: 31261630]
14. Sammani A, Jansen M, Linschoten M, Bagheri A, de Jonge N, Kirkels H, et al. UNRAVEL: big data analytics research data platform to improve care of patients with cardiomyopathies using routine electronic health records and standardised biobanking. Neth Heart J 2019 May 27;27(9):426-434 [FREE Full text] [doi: 10.1007/s12471-019-1288-4] [Medline: 31134468]
15. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. JMIR Med Inform 2016 Nov 21;4(4):e38 [FREE Full text] [doi: 10.2196/medinform.5359] [Medline: 27872036]
16. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. JMIR Med Inform 2014 Jan 17;2(1):e1 [FREE Full text] [doi: 10.2196/medinform.2913] [Medline: 25600256]
17. Li R, Niu Y, Scott SR, Zhou C, Lan L, Liang Z, et al. Using electronic medical record data for research in a health care information and management systems society (HIMSS) analytics electronic medical record adoption model (EMRAM) stage 7 hospital in Beijing: cross-sectional study. JMIR Med Inform 2021 Aug 03;9(8):e24405-e24405|p.101 [FREE Full text] [doi: 10.2196/24405] [Medline: 34342589]
18. Li P, Xie C, Pollard T, Johnson AEW, Cao D, Kang H, et al. Promoting secondary analysis of electronic medical records in China: summary of the PLAGH-MIT critical data conference and health datathon. JMIR Med Inform 2017 Nov 14;5(4):e43 [FREE Full text] [doi: 10.2196/medinform.7380] [Medline: 29138126]
19. Liu L, Liu ZY. Research on the development and application of big data in health care. Smart Healthc. 2020 2020 Aug 15;6(23):1-10. [doi: 10.19335/j.cnki.2096-1219.2020.23.001]
20. Khatri V, Brown CV. Designing data governance. Commun ACM 2010 Jan 01;53(1):148-152. [doi: 10.1145/1629175.1629210]
21. Chang Z, Chen M. Research on governance methods of health care resources in big data era. China Digit Med 2016 Sep 15;09(11):2-5. [doi: 10.3969/j.issn.1673-7571.2016.09.001]
22. Fei X, Li J, Huang Y, Wei L, Liang Z. Data governance in health care big data application. Chin J Health Inform Manag 2018 Oct 20;15(05):554-558. [doi: 10.3969/j.issn.1672-5166.2018.05.016]
23. Wang X, Xu X, Zhou G, Yang Z, Zhang Y. Research on the construction method of the tag system for health care big data. Chin J Health Inform Manag. 2021 2021 Apr 20;02(18):189-193. [doi: 10.3969/j.issn.1672-5166.2021.02.08]
24. Du Y, Gong C, Fu A, Wang D, Yin S, Zhang J. Research on Harvard datatags system and its inspiration for China. Libr J 2019 Aug 15;38(08):17-26. [doi: 10.13663/j.cnki.lj.2019.08.002]
25. Bar-Sinai M, Sweeney L, Crosas M. Datatags, data handling policy spaces and the Tags Language. 2016 Nov Presented at: IEEE Security and Privacy Workshops; May 22-26, 2016; San Jose, California p. 1-8. [doi: 10.1109/spw.2016.11]
26. Zhang N, Shi HX, Xie Q, Wang B, Zhou HW, Zhang L, et al. Ethical issues of medical data sharing under the background of big data. Chin J Inf Tradit Chin Med 2018 Jul 18;25(08):9-11. [doi: 10.3969/j.issn.1005-5304.2018.08.003]
27. Xin HY, Li P, Zhang GQ. Construction and application of medical research big data platform in hospital. Chin J Health Inform Manag 2019 Apr 20;16(02):206-209. [doi: 10.3969/j.issn.1672-5166.2019.02.019]
28. Yao AC. Protocols for secure computations. 1982 Presented at: 23rd Annual Symposium on Foundations of Computer Science; November 3-5, 1982; Chicago, Illinois, USA.
29. Stausberg J, Nasseh D, Nonnemacher M. Measuring data quality: a review of the literature between 2005 and 2013. Stud Health Technol Inform 2015;210:712-716. [doi: 10.3233/978-1-61499-512-8-712] [Medline: 25991245]
30. Weiskopf NG, Weng CH. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]

31. Meng RG, Yang Y, Zhang LX. Approaches and prospects on data quality assessment of big data in health and medicine. Chin J Health Inform Manag 2019 Dec 20;16(06):677-681. [doi: 10.1007/978-3-319-77525-8_100271]
32. Wang C, Wang Z, Wang G, Lau JY, Zhang K, Li W. COVID-19 in early 2021: current status and looking forward. Signal Transduct Target Ther 2021 Mar 08;6(1):114 [FREE Full text] [doi: 10.1038/s41392-021-00527-1] [Medline: 33686059]
33. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Jun 11;181(6):1423-1433.e11 [FREE Full text] [doi: 10.1016/j.cell.2020.04.045] [Medline: 32416069]
34. Xu X, Wang C, Guo J, Gan Y, Wang J, Bai H, et al. MSCS-DeepLN: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Med Image Anal 2020 Oct;65:101772. [doi: 10.1016/j.media.2020.101772] [Medline: 32674041]
35. Wang G, Liu X, Shen J, Wang C, Li Z, Ye L, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. Nat Biomed Eng 2021 Jun;5(6):509-521 [FREE Full text] [doi: 10.1038/s41551-021-00704-1] [Medline: 33859385]
36. Luo G, Yang Q, Chen T, Zheng T, Xie W, Sun H. An optimized two-stage cascaded deep neural network for adrenal segmentation on CT images. Comput Biol Med 2021 Sep 08;136:104749-104749. [doi: 10.1016/j.compbiomed.2021.104749] [Medline: 34388467]

Abbreviations

CT: computed tomography
GPU: graphics processing unit
WCH: West China Hospital of Sichuan University
WCH-BDP: West China Hospital of Sichuan University big data platform

Edited by C Lovis; submitted 16.01.22; peer-reviewed by J Lei, W Meng; comments to author 04.02.22; revised version received 17.02.22; accepted 25.02.22; published 13.04.22.

Please cite as:
Wang M, Li S, Zheng T, Li N, Shi Q, Zhuo X, Ding R, Huang Y
Big Data Health Care Platform With Multisource Heterogeneous Data Integration and Massive High-Dimensional Data Governance for Large Hospitals: Design, Development, and Application
JMIR Med Inform 2022;10(4):e36481
URL: https://medinform.jmir.org/2022/4/e36481
doi:10.2196/36481
PMID:35416792

Original Paper

Automating Large-scale Health Care Service Feedback Analysis: Sentiment Analysis and Topic Modeling Study

George Alexander[1*], MSci; Mohammed Bahja[1*], PhD; Gibran Farook Butt[1*], MD

The School of Computer Science, University of Birmingham, Birmingham, United Kingdom

[*]all authors contributed equally

Corresponding Author:
George Alexander, MSci
The School of Computer Science
University of Birmingham
The School of Computer Science, The University of Birmingham
Edgbaston
Birmingham, B15 2TT
United Kingdom
Phone: 44 0121 414 3344
Email: gxa548@alumni.bham.ac.uk

Abstract

Background: Obtaining patient feedback is an essential mechanism for health care service providers to assess their quality and effectiveness. Unlike assessments of clinical outcomes, feedback from patients offers insights into their lived experiences. The Department of Health and Social Care in England via National Health Service Digital operates a patient feedback web service through which patients can leave feedback of their experiences in structured and free-text report forms. Free-text feedback, compared with structured questionnaires, may be less biased by the feedback collector and, thus, more representative; however, it is harder to analyze in large quantities and challenging to derive meaningful, quantitative outcomes.

Objective: The aim of this study is to build a novel data analysis and interactive visualization pipeline accessible through an interactive web application to facilitate the interrogation of and provide unique insights into National Health Service patient feedback.

Methods: This study details the development of a text analysis tool that uses contemporary natural language processing and machine learning models to analyze free-text clinical service reviews to develop a robust classification model and interactive visualization web application. The methodology is based on the design science research paradigm and was conducted in three iterations: a sentiment analysis of the patient feedback corpus in the first iteration, topic modeling (unigram and bigram)–based analysis for topic identification in the second iteration, and nested topic modeling in the third iteration that combines sentiment analysis and topic modeling methods. An interactive data visualization web application for use by the general public was then created, presenting the data on a geographic representation of the country, making it easily accessible.

Results: Of the 11,103 possible clinical services that could be reviewed across England, 2030 (18.28%) different services received a combined total of 51,845 reviews between October 1, 2017, and September 30, 2019. Dominant topics were identified for the entire corpus followed by negative- and positive-sentiment topics in turn. Reviews containing high- and low-sentiment topics occurred more frequently than reviews containing less polarized topics. Time-series analysis identified trends in topic and sentiment occurrence frequency across the study period.

Conclusions: Using contemporary natural language processing techniques, unstructured text data were effectively characterized for further analysis and visualization. An efficient pipeline was successfully combined with a web application, making automated analysis and dissemination of large volumes of information accessible. This study represents a significant step in efforts to generate and visualize useful, actionable, and unique information from free-text patient reviews.

(JMIR Med Inform 2022;10(4):e29385) doi:10.2196/29385

KEYWORDS

natural language processing; topic modeling; National Health Service; latent Dirichlet allocation; reviews; patient feedback; automated solutions; large-scale health service; free-text; unstructured data

Introduction

Background

Patient experience is described by the Beryl Institute as "the sum of all interactions, shaped by an organisation's culture, that influence patient perceptions across the continuum of care" [1]. It is a vital consideration of the health service provider's planning strategy to reflect patient engagement and service quality [2]. It is also a contributing factor to patient engagement, which is key to delivering effective and efficient care [3-6]. The *Patient experience improvement framework* of the National Health Service (NHS) defines several quality indicators, among which patient feedback is a priority [7]. Thus, to deliver truly patient-centered care, the patient experience must be a central consideration [8], and health care providers must have mechanisms in place through which the patient experience can be understood.

Over the past 2 decades, there has been a greater emphasis on using patient feedback to inform and improve service delivery, largely in the United Kingdom, Europe, and the United States [9,10]. The way feedback is obtained can vary greatly, ranging from individual interviews and focus groups to official complaints as well as surveys conducted through various media (postal or web-based). Surveys and similar quantitative methodologies generate measurable results that can be used for benchmarking and comparisons over time or between subjects. Although they can help identify some problem areas in services, they can lack the specificity required to drive change [10-12]. Feedback mechanisms that allow in-depth ideas to be shared, such as patient forums, can help generate detailed patient experience insights [10-12].

The NHS website allows users to anonymously rate and share their experience in a public forum [13]. All NHS-provided services across England can be found on the site, where users can leave a free-text comment and give an overall *star* rating out of 6. These publicly available reviews are invaluable insights into the work of the NHS for the service providers themselves, national bodies such as NHS England, and the Care Quality Commission as well as patients deliberating on which services to use. This is a source of vast amounts of rich data, which has the potential to significantly influence the quality of services nationwide as well as policy regarding the NHS. Patient feedback in free-text form is typically hard to analyze on a large scale, which is why standardized scales are more frequently used to generate numerical measures for comparisons [14]. The difficulty from the patients' perspective lies in the accessibility to the data, which is limited to scrolling through individual responses in a particular service.

This type of data lends itself well to analysis using computed natural language processing (NLP) techniques, enabling high-volume automated analysis of text information. In their seminal work, Greaves et al [15] reported on the utility of this NHS web-based feedback data to gain insights into the health care service while allaying concerns about the risk of unsolicited reviews creating biased feedback. Greaves also reported that, with regard to the accuracy of the feedback about a given clinical service, web-based feedback was comparable with conventional surveys of patient experience [15].

The advent of machine learning and the development of sophisticated NLP algorithms have significantly advanced the analysis of text corpora. A significant amount of research has focused on applying advanced NLP methods to web-based reviews. Web-based reviews provide an opportunity to explore free-text corpora that do not usually adhere to a structure or format. The free text in reviews, such as the patient experience, makes the process of automated analysis of the review challenging when compared with closed questions with an expected text input. As web-based patient feedback is extensive, the traditional text analysis methods may provide limited capabilities for analysis. The latest machine learning– and artificial intelligence–based NLP methods have been well explored for analyzing large review data sets, especially for analyzing the user experience. The latest NLP methods provide capabilities to classify the reviews as positive or negative with high accuracy. Identifying the underlying themes and topics in the user experience allows us to understand the frequently reported service areas in user feedback.

Objective

The aim of the study presented in this paper is to provide an automated solution for the large-scale analysis of patient feedback on health care service providers. This study achieves this by exploring NLP techniques, including sentiment analysis and topic modeling. The objective of this study is also to present an interactive interface that provides stakeholders with a portal to analyze and identify outcomes of the patient feedback analysis. This work builds on previous work in the field [16,17] and presents the design and implementation of an unsupervised machine learning NLP model combined with an interactive interface to produce a user-friendly web application that allows for the exploration of service reviews across England.

Methods

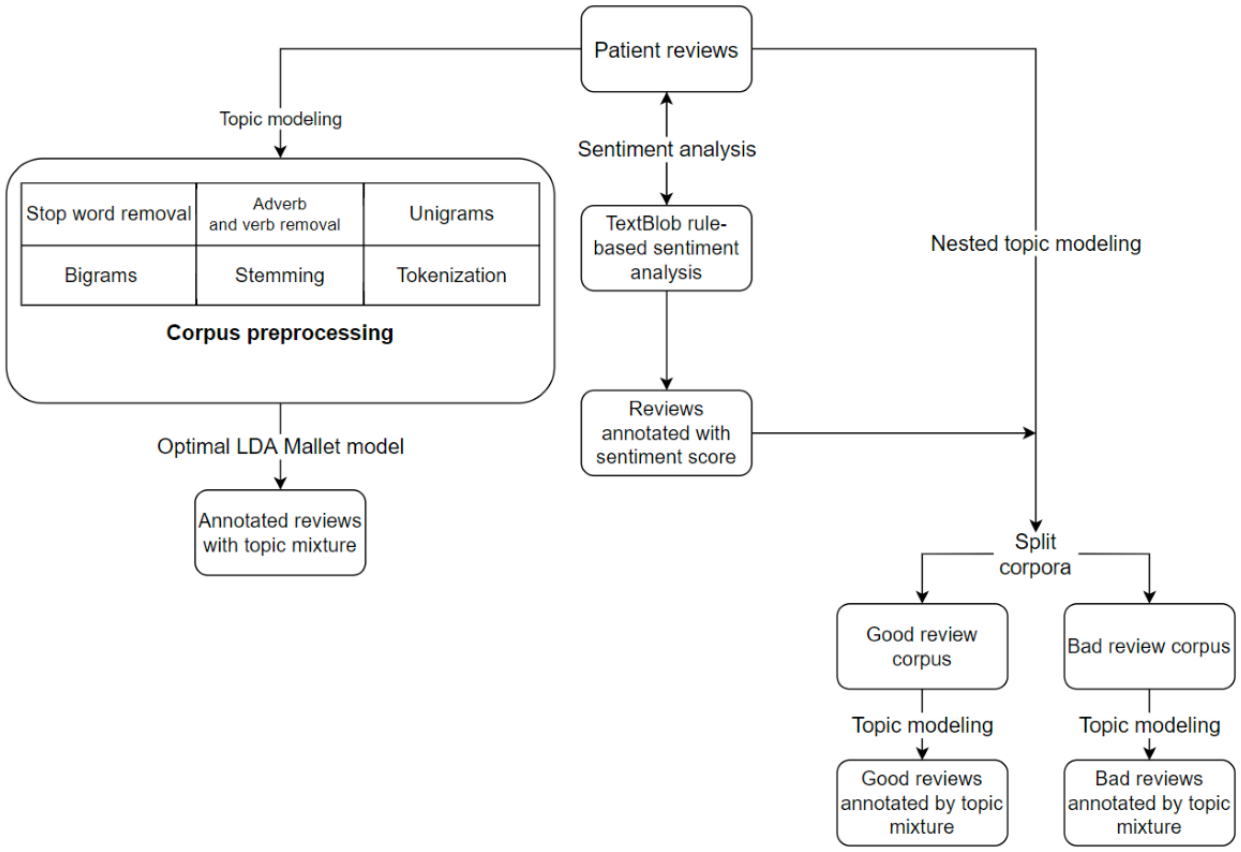
Overview

For this study, the design science research (DSR) methodology was used. The DSR paradigm is a widely popular research approach in information systems research. It is referred to as a problem-solving paradigm as it aims to build *artifacts* that are aimed at addressing a problem. The artifacts address the problems or enhance existing solutions and are important tools for arriving at research outcomes and reviewing them to decide how the adopted artifact can be further used [18]. The DSR process follows a systematic procedure in which the artifacts are developed through the systematic creation, capturing, and communication of knowledge from the design process. DSR uses an iterative process whereby the artifacts are reconstructed at each iteration and, thus, can be described as a continuous learning process that enhances the artifact quality incrementally [19]. Further details on the DSR methodology in this study can be found in our previous study [17].

Figure 1 illustrates the methodology carried out for this research. The patient review corpus was subjected to three different iterations of NLP analysis: sentiment analysis, topic modeling,

and nested topic modeling. The first iteration, sentiment analysis, enabled the automated analysis of the patient reviews and identified the sentiment of the reviews as either positive or negative. The second iteration, topic modeling analysis, applied the unigram and bigram topic modeling methods to annotate reviews with a group of words that reflected a theme or topic. A single review might have one or more topics. In the third iteration of the study, a nested topic modeling approach was applied. Nested topic modeling analysis combines sentiment analysis and topic modeling methods. The patient reviews were annotated with their associated sentiment score and then split into the *good* corpus and *bad* corpus based on the associated positive or negative sentiment. The *good* corpus and *bad* corpus were analyzed with topic modeling to identify topics within the reviews.

Figure 1. The research methodology followed for the analysis of the patient feedback corpus. LDA: latent Dirichlet allocation.



NHS Patient Feedback

The NHS website includes a platform where patients provide both ratings and reviews for a particular NHS service. The NHS website rating system provides an outline of patient experience; the rating is an optional feature that is collected for a specific set of parameters such as *cleanliness* and *dealt with dignity*, among others.

Patients can provide feedback about NHS hospitals in 3 main sections. First, they are asked to rate, on a scale of 1 to 6, *how likely they are to recommend the particular hospital to family and friends?* This is the central question on the NHS website in that the ratings provided are used to calculate the overall rating for a given hospital. The rating for this question provides a quick and easy indicator of a hospital's performance in providing patient care. However, the rating is single and straightforward, and it is insufficient to obtain a detailed understanding of hospital performance. For a more detailed understanding of patient feedback, the NHS website includes questions where the patients are asked to provide ratings on five parameters: cleanliness, staff co-operation, dignity and respect, involvement in decisions, and same-sex accommodation (out of 6 stars). The website-allocated ratings for these 5 parameters are optional to the users. Finally, there is an optional free-text review of a maximum of 3000 characters. These reviews follow the NHS comment policy and are moderated. The moderators remove any personal information and ensure that the reviews do not cause any legal issues such as defamation [13]. Responses to the 5 parameters and free-text data provide an opportunity for a granular assessment and understanding of patient feedback for a hospital.

All available reviews between October 1, 2017, and September 30, 2019, were collected from the NHS website.

The NHS platform provides an application programming interface (API) that allows access to the patient ratings and reviews [20]. A custom web scraper was built for the project using the NHS API to collect the patient ratings and reviews. The NHS platform provides the data with the standard license terms that cover the requirements of the General Data Protection Regulations [21].

Data Preprocessing

The data set underwent a few processing steps where only columns from the data set that were relevant for this study were selected. Specifically, for parsimony purposes, only relevant data fields were extracted and relabeled into *Date*, *Comment*, and *Label* columns in the database. The *Date* and *Comment* columns referred to the posted date and the content of the participant's comment, and the *Label* column was used to hold the sentiment of the comment, classified as either positive or negative.

Data were organized by posting date [22,23] and partitioned into training, test, and validation data sets. In the training data set, observations were labeled according to the sentiment inferred from the comments. The test data set was used as an input to derive patterns from the training data set using text-mining models. In the validation data set, the values in the *Label* column were not defined.

Within the data set, the ratings given by the participants for the following question—*How likely are you to recommend this hospital to friends and family if they needed similar care or treatment?*—were used as the actual data against which the performance of the text-mining model (ie, in predicting the patient feedback) was tested. Owing to the skewed distribution of the numerical responses and limitations of the machine learning methods, to reduce complexity in the modeling procedure, the continuous-scale patient feedback ratings were discretized. Following discretization, rating scores of 1 and 2 were categorized as negative, and those of 5 and 6 were categorized as positive. Rating scores of 3 and 4 were discarded as they were deemed neutral ratings that did not portray polarization. This categorization served as a binary sentiment label for which each text-mining model was trained and assessed.

Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, refers to the computational study of people's opinions, sentiments, attitudes, and emotions toward an entity. The entity can be another individual or a public figure, a product such as an electronic device, or service providers such as restaurants and hospitals. The identification of sentiment is performed based on the presence of words or phrases that are likely to refer to an opinion, sentiment, or emotion. If the sentence is identified as having a sentiment or opinion, it is then subjected to the feature selection process. During this process, the identified sentiment is associated with the feature that is being discussed. Finally, the sentiment is categorized into a chosen classification type. For instance, the sentiment can be classified into a binary, such as positive or negative. The sentiment analysis is associated with a score.

The fine-grained sentiment score detects polarity within a text; in this case, whether the review is a positive or negative opinion. There are several approaches to sentiment analysis, including the strength of association, naïve Bayes (NB), and the support vector model. NB-based sentiment analysis models are popular and widely used. The NB classifier is a probabilistic classifier, which uses a mixture of models for classification and is widely popular for sentiment classification. Given a document, and based on the distribution of words in the document, the NB approach computes the probability of a document belonging to a class. This model calculates boundaries according to the distribution of the words across the labels while at the same time considering the joint probability of the words occurring independently together. Specifically, NB considers each word independently of one another and then tries to estimate the posterior distribution of a review being positive or negative according to the joint distribution of the words in the review. The probability is computed using the Bayes theorem to predict that a given word belongs to a specific sentiment.

One of the popular implementations of sentiment analysis based on NB is the TextBlob rule-based sentiment analysis [24], which was adopted in our study. The TextBlob approach allows for the performance of different NLP tasks, including part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and translation. The TextBlob approach is suitable in the current version of our study when compared with advanced machine learning–based approaches because of the relatively smaller size of the data set that is used. A machine learning–based method inherently requires large, labeled data sets for training and testing that could be prohibitive and expensive [25].

The TextBlob implementation was used to analyze each of the 51,845 reviews to determine their sentiment value. This produced a score for each review between −1 and 1, where 1 represents a wholly positive sentiment, −1 represents an entirely negative sentiment, and 0 represents a neutral sentiment. This provided a method for evaluating whether reviews were *good* or *bad* more quickly and consistently than a human-based process.

Topic Modeling

Topic modeling is an unsupervised NLP approach where unlabeled documents are used to create a set of topics represented by a list of words that frequently occur in each topic. There are several topic modeling approaches, and most use dimensionality-reducing techniques with the goal of representing a document using fewer words. Some of the most popular topic modeling approaches are probabilistic latent semantic indexing (LSI), latent Dirichlet allocation (LDA), and correlated topic modeling (CTM).

The LSI approach uses linear algebraic approaches such as singular value decomposition and *bags of words* to represent documents. It aims to extract words that carry similar meanings (ie, it uses synonyms and polysemy for topic identification [26]). The LSI approach assumes that each document has multiple topics and that the probability of each consists of a weight for a given document. On the basis of this assumption, the topics in a document are identified.

A disadvantage of the LSI approach is that the number of parameters in the model increases as the volume of data increases, and this could lead to overfitting problems. Furthermore, when the LSI model is used on documents that were not part of the training data set, the topic probabilities have to be assigned again [27].

The CTM approach helps in identifying the correlation between a specific topic and others. The correlation information might help the users in identifying links or associations between a specific topic from a database and other similar topics. The CTM approach uses a logistic normal distribution to identify topics from documents. A covariance matrix used for parameterizing the distribution is then used to identify the correlation between the topics. A topic graph is subsequently drawn, in which the topics are represented by nodes and their correlations with other topics are depicted [28]. The correlation approach provides more information to the user and, thus, enables better interpretation of the information. The CTM approach achieves a higher predictive likelihood than the LSI approach [29].

The LDA approach also works under the assumption of LSI (ie, that each topic is a distribution of words and each document has a certain distribution of topics). However, this assumption is extended by using a hidden variable model of documents that consists of hidden random variables with which the observed data interact [30]. In LDA, the hidden variables are the topics and how the document exhibits them, and the observed data are the words. The learned or posterior distribution of the hidden variables for the given documents determines their topical composition. Furthermore, the LDA approach uses the Dirichlet distribution to define the distribution of topics in a document [31].

An advantage of the LDA approach is that the statistical assumptions it makes for topic modeling enable it to uncover sophisticated structures in the texts. For instance, the *bag of words* assumption used in the LDA approach makes it invariant to the order of words in the document. Furthermore, the order of documents in the corpus is also not a criterion for the LDA approach to extract topics from the document. This might not be suitable if the patient's experience needs to be analyzed longitudinally (ie, over a period). However, in this study, as the patient experience analysis did not consider the time factor, the LDA method suited its aims.

We used the LDA topic modeling approach to categorize each review into computer-generated topics. LDA initially assumes the number of topics and attempts to calculate topics that best represent the documents. It does this by calculating the probability estimate of a word for a given topic as well as the probability of a topic for a given document [32].

We used an LDA implementation called LDA Mallet owing to its use of a more precise sampling method called Gibbs sampling [33]. Data were preprocessed, including the removal of stop words, verbs, and adverbs as well as lemmatization, formation of bigrams, and conversion of the corpus into the bag-of-words format. Bigrams were generated using the Gensim library (RARE Technologies, Ltd), which automatically detects common phrases [34]. LDA Mallet has 2 parameters, a number of topics, and a hyperparameter α. These parameters were optimized through a series of experiments, with the number of topics ranging from 5 to 25 and α ranging from .01 to .99. Each test was measured using the coherence score [35] as well as using human judgment to determine the validity of the generated topics. The highest-rated LDA model was then used to determine the topic mixture of each review.

Reviews were then labeled according to the dominant topic. The dominant topic was defined by the LDA model, predicting the percentage contribution of that topic to be ≥50%. The reviews that did not have any topic that contributed ≥50% were not included in the following analysis.

Nested Topic Modeling

The corpus was divided into two subcorpora, the first one being negative-sentiment–scoring reviews and the second one being any positive-sentiment–scoring reviews. Similar to the topic model for the entire corpus, we then performed experiments to determine the optimal parameters for both corpora. Using these parameters, we produced two models, one showing the topics generated from negative-sentiment reviews and the other showing topics generated from positive-sentiment reviews. Applying these new LDA models to their respective corpora produced a topic mixture for each entry of their respective corpora.

The nested topic modeling approach enables the identification of the rationale behind a particular sentiment of the patient in each comment or review. The intent is to find out why the patient was happy or unhappy about a particular topic in each comment. The problem being addressed in this iteration of the study was to find the possible reason behind a patient's sentiment for a particular topic in each comment.

Visualization

Visualizing the results from the NLP methods relied on a Microsoft Azure Cloud Service [36] to host the SQL server, web functions (API), and the web application. The SQL server stored each review and service as well as the results from both the sentiment analysis and topic modeling. The web functions acted as an API to allow for a Representational State Transfer and secure connection to the database. Finally, the cloud service hosted the web application, which used NodeJS (OpenJS Foundation) [37] as the back-end framework and VueJS [38] as the front-end framework. The web application also used packages such as Google Maps [39] and VueChartJS [40]. Both of these packages ensured that the results were shown in a logical, effective, and efficient way.

Results

NHS Patient Feedback

NHS England is segmented into seven geographical regions with teams supporting the delivery of care locally: London, Midlands, North East and Yorkshire, North West, East of England, South East, and South West. Of the 11,103 possible services across these regions, 2030 (18.28%) services received a combined total of 51,845 reviews between October 1, 2017, and October 31, 2019. Among the reviewed services, the mean number of reviews per service was 26 (SD 60.3), and the highest number of reviews for a single service was 550 for Lincoln County Hospital. During the study period, the mean number of reviews per month across England was 2028 (SD 449.1). The

number of reviews per month declined from 2625 in October 2017 to 1611 in September 2019.

The number of services per 10,000 population was similar around England, with the lowest being in London, which also has the highest-density population (Table 1). Across England, 18.11% (2011/11,103) of the services received a review during the study period, with the fewest services reviewed being in the North East and Yorkshire and the most reviewed being in London. The number of reviews per 10,000 population across England was similar, with the fewest being in the South East region and the highest in the East of England (Table 1).

Table 1. National Health Service England regions, services, and reviews.

Characteristic	England	East of England	London	Midlands	North East and Yorkshire	North West	South East	South West
Population (million)[a]	53.8	4.5	8.2	10.1	7.9	7.1	8.6	5.3
Services, n (%)	11,103 (100)	1095 (9.9)	1343 (12.1)	2123 (19.1)	1865 (16.8)	1572 (14.2)	1929 (17.4)	1176 (10.6)
Reviewed services, n (%)	2011 (18.1)	221 (20.2)	309 (23)	369 (17.4)	299 (16)	269 (17.1)	323 (16.7)	221 (18.8)
Unreviewed services, n (%)	9092 (81.9)	874 (79.8)	1034 (77)	1754 (82.6)	1566 (84)	1303 (82.9)	1606 (83.3)	955 (81.2)
Total reviews, n (%)	50,707 (100)	5517 (10.9)	8337 (16.4)	10,047 (19.8)	7212 (14.2)	6896 (13.6)	7709 (15.2)	4989 (9.8)
Reviews per 10,000	9.4	12.3	10.2	9.9	9.1	9.7	9.0	9.4
Services per 10,000	2.1	2.4	1.6	2.1	2.4	2.2	2.2	2.2

[a]Population data from the Office for National Statistics Census, 2011.

Sentiment Analysis

To explore the opinions held within the reviews, the sentiment was analyzed for the entire corpus and for individual reviews. The analysis generated a score for how positive (eg, happy or pleased) or negative (eg, unhappy or disappointed) the sentiment was between −1 (most negative) and 1 (most positive). Examples of reviews and their corresponding score can be found in Table 2. The average sentiment of all reviews across the study period did not demonstrate any significant changes (Figure 2A). A comparison of sentiment by season revealed that there were no significant differences among spring (March 20 to June 21), summer (June 21 to September 22), and autumn (September 22 to December 21); however, a statistically significant decrease occurred in winter (December 21 to March 20; mean sentiment 0.178, SD 0.21) compared with summer (mean sentiment 0.185, SD 0.21; P=.02).

Table 3 reports the sentiment scores for each of the topics in the corpus. Topics that are innately associated with positive sentiments, such as good experience and good staff, have higher sentiment scores than inherently negative topics such as rude staff. The distribution of sentiment and topic frequency demonstrates a tendency in the most frequently mentioned topics to be the most polarized, appearing as a u-shaped curve (Figure 2B). The topic sentiment scores over time appeared to vary around their overall corpus sentiment score and did not significantly change throughout the study period. Consultancy appeared to vary most significantly across the study period. Mental health demonstrated a downward trend toward the latter part of the study period. Good experience, good staff, and operations and surgery had consistently higher sentiment scores across the study period compared with the other topics (Figure 2C).

Table 2. Sentiment scale examples.

Sentiment score	Review
−1.0	“Awful treatment in the SAU[a] area. Avoid using if you can!”
−0.5	“I think there a load of rubbish whenever you ring them they dont answer tried 8 times today what a joke.”
0.0	“Have been attending clinic regularly. Always have eye problems after drops. Despite several attempts to bring this to staff attention it has been dealt with in a dismissive manner. GP[b] attendance has been necessary. Phone always engaged or left ringing.”
0.1	“All midwives very helpful. An improvement could be that proper beds are provided for partners.”
0.2	“I felt very respected and would like to say thank you.”
0.5	“Excellent result from the procedure performed from the team, my sight is back to what I want. As always (department name) do the business.”
1.0	“The care of all the staff was excellent; and some of them deserved an MBE[c].”

[a]SAU: surgical assessment unit.

[b]GP: general practitioner.

[c]MBE: Member of the Order of the British Empire.

Figure 2. A collection of relationships within the data set. A: sentiment score over time; B: Relationship between topic frequency and sentiment score over time; C: Relationship between topic frequency and sentiment. A&E: accident and emergency.

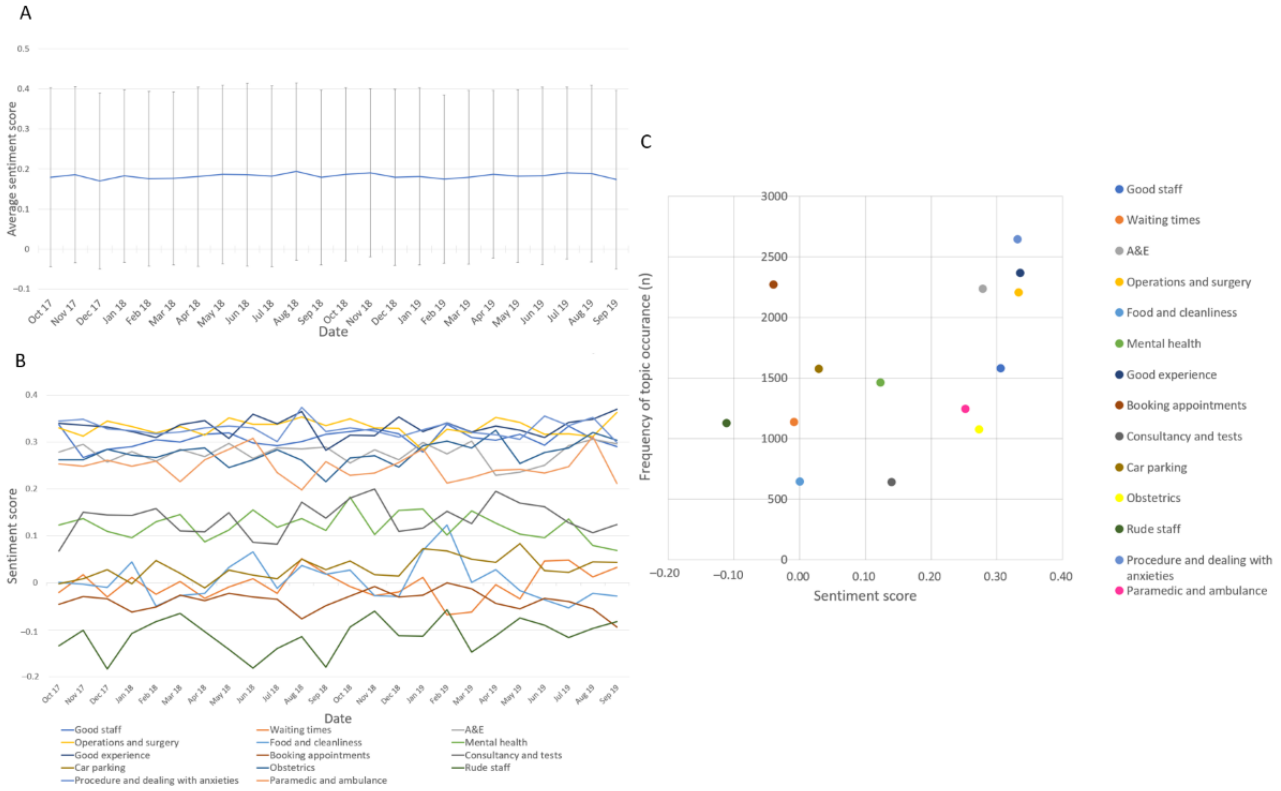


Table 3. Latent Dirichlet allocation–generated topics with their sentiment score (refer to Multimedia Appendix 1 for keywords generated for each cluster).

ID	Topic	Classified reviews, n (%)	Sentiment score
1	Good staff	1579 (7.1)	0.305795
2	Waiting times	1136 (5.1)	−0.006903
3	A&E[a]	2238 (10.1)	0.277627
4	Operations and surgery	2205 (9.9)	0.331010
5	Food and cleanliness	647 (2.9)	0.002680
6	Mental health	1463 (6.6)	0.122655
7	Good experience	2369 (10.7)	0.333013
8	Booking appointments	2272 (10.2)	−0.038362
9	Consultancy and tests	640 (2.9)	0.139792
10	Car parking	1577 (7.1)	0.028805
11	Obstetrics	1075 (4.8)	0.272878
12	Rude staff	1128 (5.1)	−0.110659
13	Procedure and dealing with anxieties	2646 (11.9)	0.330488
14	Paramedic and ambulance	1246 (5.6)	0.250670
N/A[b]	Total	22,221 (100)	0.182695

[a]A&E: accident and emergency.

[b]N/A: not applicable.

Topic Modeling

A total of 14 clusters were identified from the entire corpus, from which themes were derived manually (Table 3). Reviews were then classified according to their dominant topic (ie, the topic to which the review had a >50% probability of belonging as identified by the LDA model). This threshold was selected to reduce the confounding effect of sentiment analysis by

co-occurring opposing sentiments that may be encountered in reviews that contained multiple topics. Associations between topic frequency and geographic distribution and their changes over time were then characterized.

In total, 22,221 reviews were classified with a dominant topic, as shown in Table 3, and the topics identified were reviewed and labeled by GFB, who is a medical doctor in the NHS. The most frequent topics identified were *paramedic and ambulance* (topic 14; 1246/22,221, 5.61%), *booking appointment* (topic 8; 2272/22,221, 10.22%), *good experience* (topic 7; 2369/22,221, 10.66%), *operations and surgery* (topic 4; 2205/22,221, 9.92%), and *A&E* (topic 3; 2238/22,221, 10.07%). These topics comprised 46.49% (10,330/22,221) of the labeled reviews. Most topics occurred at a steady rate across the study period. No patterns in that variation in topic frequency were identified; however, *procedure and dealing with anxieties* increased, whereas *obstetrics* decreased in frequency across the study period (Figure 3). Most topics were of similar proportions across all regions in England; however, *waiting time* and *A&E* were proportionately greater in London and the South West, respectively (Figure 4).

Figure 3. Topics over time. A&E: accident and emergency.

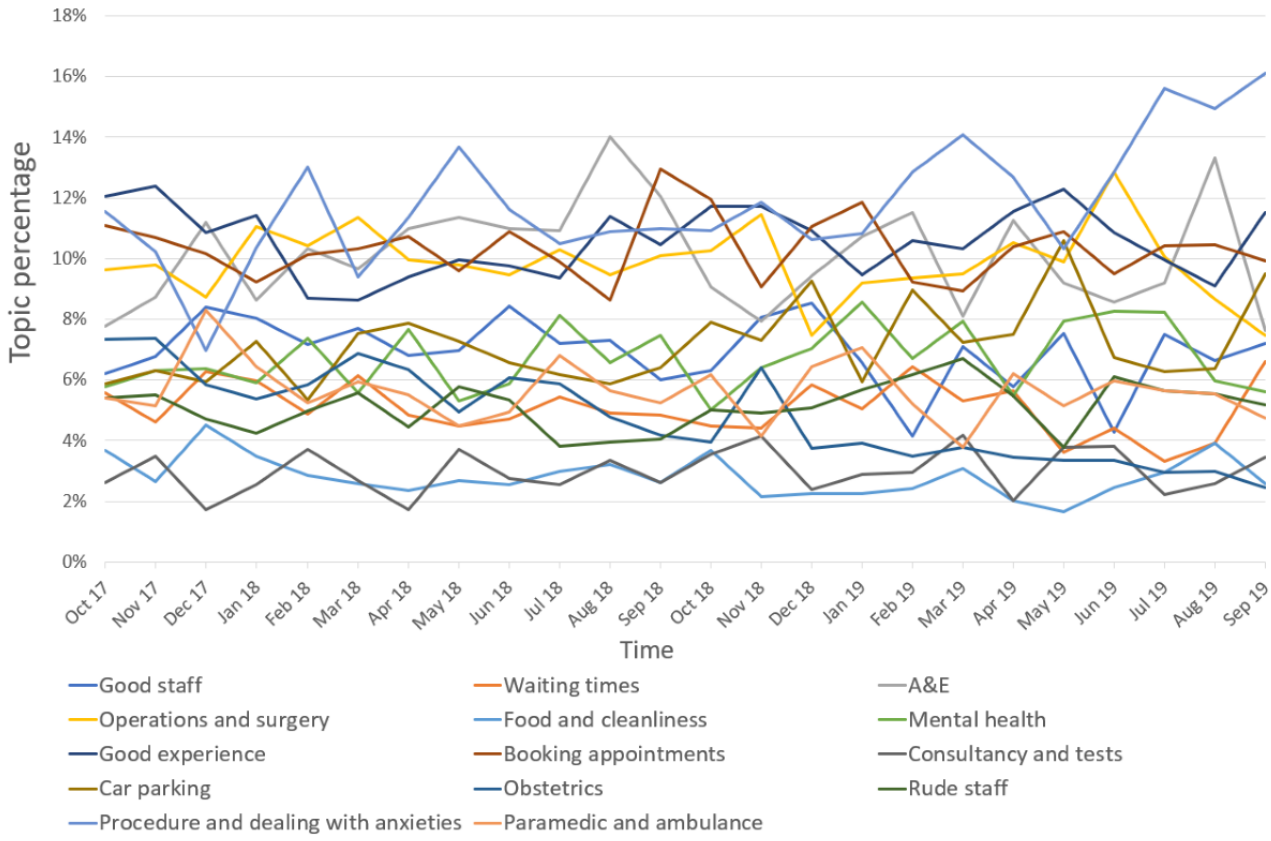
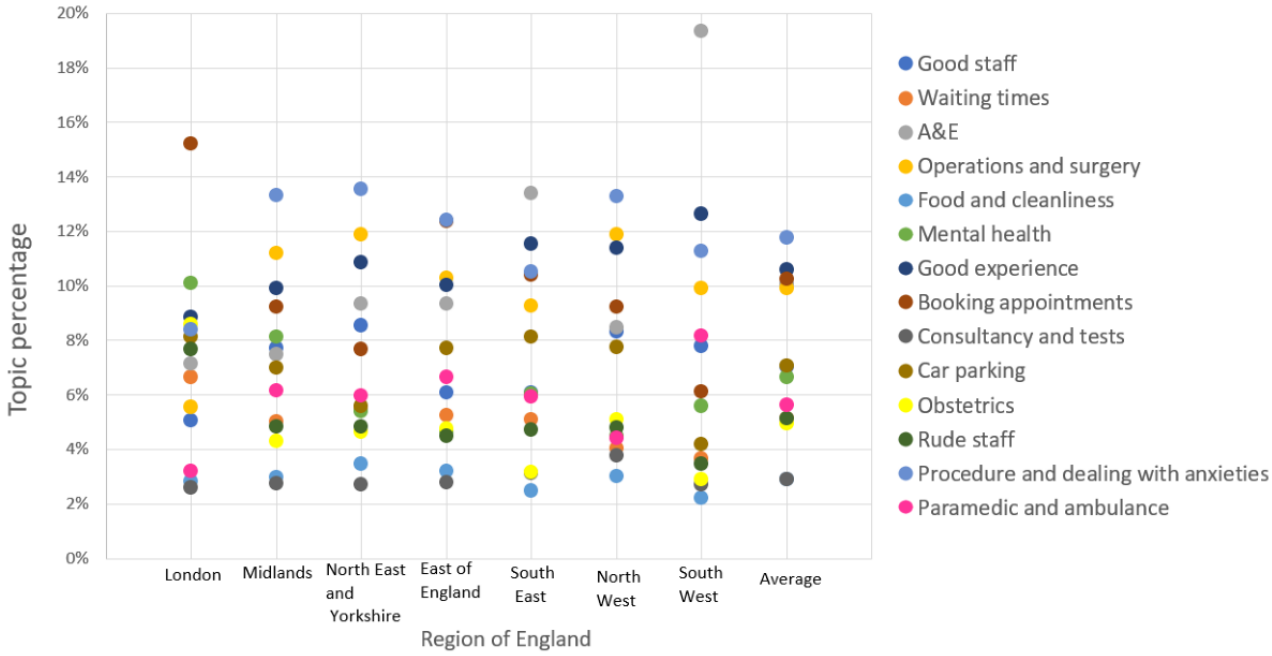


Figure 4. Topic frequency across different regions. A&E: accident and emergency.



Nested Topic Modeling

Using sentiment scoring, positive-scoring reviews (≥0.2) and negative-scoring reviews (≤0.2) were separated into 2 smaller corpora to undergo topic modeling to obtain new sets of positive and negative topics. The optimal number of topics was decided by creating a model for each number of topics in the range of 5 to 20 and using the coherence score as well as human intuition to determine the best topic model. Positive and negative clusters were identified, and the topics were labeled manually (Tables 4 and 5). The sentiments of both the positive and negative topics demonstrated typical undulations over time, remaining around their average. Regarding positive topics, *admissions* and *surgery* demonstrated rapid increases and decreases back to their average in 2019 and, of the negative topics, *rude*, *booking appointment*, and *food and cleanliness* scored the lowest sentiment.

Table 4. Negative-sentiment topics (any review with a sentiment score <−0.2; refer to Multimedia Appendix 2 for keywords generated for each cluster; n=863)

ID	Human-generated name	Reviews, n (%)
1	Mental health	77 (8.9)
2	Care	58 (6.7)
3	Rudeness	65 (7.5)
4	Children	31 (3.6)
5	Pain management	81 (9.4)
6	Waiting for appointment	107 (12.4)
7	Phone	68 (7.9)
8	Cleanliness	49 (5.7)
9	Care	48 (5.6)
10	Booking appointment	159 (18.4)
11	GP[a]	31 (3.6)
12	Results	89 (10.3)

[a]GP: general practitioner.

Table 5. Positive-sentiment topics (any review with a sentiment score >+0.2; refer to Multimedia Appendix 3 for keywords generated for each cluster; n=917)

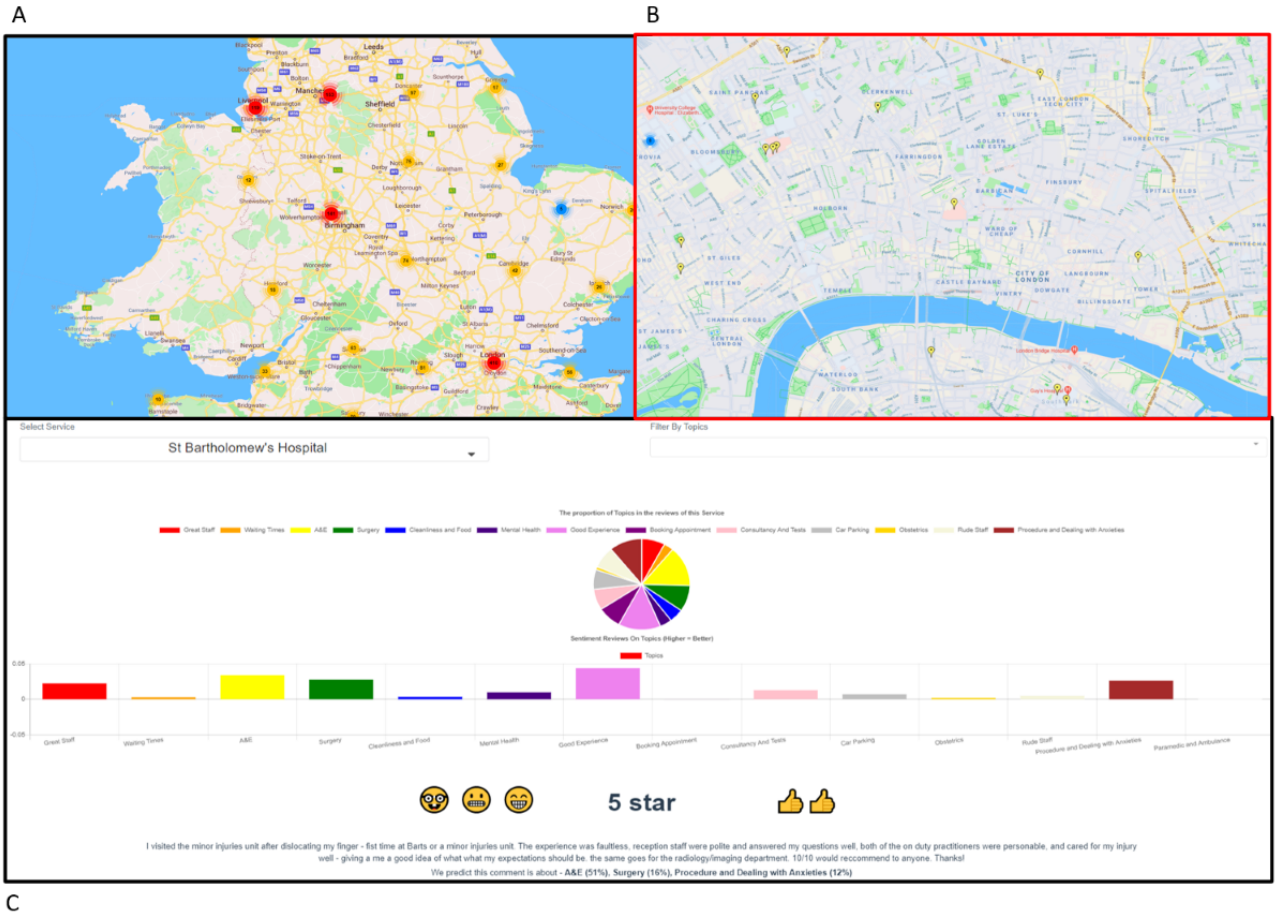
ID	Human-generated name	Reviews, n (%)
1	General care	119 (13)
2	A&E[a]	104 (11.3)
3	Admissions	57 (6.2)
4	Service	93 (10.1)
5	Pediatrics	156 (17)
6	Appointment and consultation	177 (19.3)
7	Dealing with anxieties	74 (8.1)
8	Surgery	137 (14.9)

[a]A&E: accident and emergency.

Visualization

The NLP analytics were then incorporated into a user-friendly web-based interface that enables the exploration of the data through the graphical user interface (Figure 5). Users can navigate around the country using the Google Maps–based map, which displays color-coded pins that aggregate analytics for areas (Figure 5A). The color of the pins represents the average sentiment score for that service. As the user zooms into an area, the aggregated pins are divided into color-coded pins for individual services (Figure 5B). Clicking on a service reveals a short overview of the service, displaying the average sentiment, the emotions, and the proportion of each topic derived from the reviews of that service (Figure 5B). The emotions were derived from an emotional analysis algorithm applied to each review and accumulated to find the most common emotion. This improves usability as it increases its appeal to the users.

Figure 5. Screenshots of the web-based interface user journey. A: Visualizing NHS services on a map of the United Kingdom; B: A closer view of London and its NHS services; C: The visualization of the sentiment, emotion and topic model analysis. A&E: accident and emergency; NHS: National Health Service.



Discussion

Principal Findings

The outcomes of the study presented in this paper demonstrate that NLP-based analysis using sentiment analysis and topic modeling can be used to develop an automated solution for analyzing patient feedback and reviewing the performance of a health care center. The results of the sentiment analysis showed that patient feedback can be accurately classified into positive or negative sentiments and the topic modeling approach can be used to identify topics in a patient review. Furthermore, the identified topics were associated with a sentiment based on the results of the sentiment analysis. This study also presented an interface for stakeholders to view and interact with the outcomes of the patient feedback analysis.

In this study, close to 52,000 reviews from 2030 services were considered for analysis over a period of 2 years. The average number of reviews per service was 26 (SD 60.3). This is indicative of a generally low level of patient engagement with providing service feedback. Furthermore, the number of reviews over time shows that patient review activities have seen a consistent downward trend.

A larger corpus of patient reviews can significantly contribute to building analytical models with capabilities to perform more granular analyses such as individual service performance, areas of concern for a selected service, and similar fine-grained analyses. They improve the communication channel between patients and services, and incentives for patient engagement are necessary for increasing the average number of patient reviews and contributing to the development of improved assessment tools.

The identification of a topic based on topic modeling is a useful tool as it helps in understanding the areas of performance and areas of concern for the NHS services. In this study, the most frequent topic was *procedure and dealing with anxieties* (2646/22,221, 11.91%), implying that undertaking a procedure might be an anxious process for patients, and the average sentiment score of 0.33 (SD 0.17; positive score) indicates that the patients are generally happy with the service and care provided by the NHS services.

On the contrary, another frequently reviewed topic was *booking appointments* (2272/22,221, 10.22%; however, the associated average sentiment score of −0.03 (SD 0.18) implies that the patients might be unhappy with the current appointment booking service. Therefore, topic identification, along with the associated sentiment score, allows policy makers to use the learnings from areas of success and potentially apply them to areas of concern to improve patient satisfaction with the NHS services.

The study presented demonstrates a multifactorial analysis that can be performed using topic modeling and sentiment analysis approaches. A large corpus helps in achieving temporal analysis of patient feedback, as demonstrated in Figures 2A-C. Such time-based analysis could shed light on identifying the point when the decline in an area of service occurred and the duration for which the service performed well or poorly. For instance, if the appointment booking service malfunctions frequently, and assuming patients review the appointment service, a temporal analysis could help understand the time and duration for which the service malfunctions and help the service providers diagnose the issue.

Figure 4 illustrates the topic frequency distribution for different regions of NHS centers. It can be observed that topic frequency largely varies from region to region. However, there are some commonalities in the frequently reviewed topics. For instance, the topic of *A&E* was most frequently reviewed across the regions. *Procedure and dealing with anxieties* was another frequently reviewed topic across regions. Region-based, frequently reviewed topics are a piece of beneficial information for policy makers to gain insights into the weak areas of NHS services for each region and, in general, across all regions.

The combined approach of sentiment analysis of topics identified from topic modeling methods helps collect insightful information about health care services. In our study, the sentiment classification of each topic helped in assessing the public's perception of the NHS services for a given topic, thus reflecting the service quality. With a large corpus of reviews, the sentiment analysis of identified topics over time can potentially explore the links between temporal factors, such as seasons, and patient experience for each topic. The results of this study demonstrate a decrease in sentiment during the winter. A more extensive and diverse data set of patient reviews has the potential to extract links between seasons and specific topics for further causal inferences to be made. For instance, a variation in sentiment scores for a topic across seasons can be analyzed for external causal factors such as influenza outbreaks, pandemics, staff shortages, technological hindrances, political changes, and similar outside factors.

We note that there are limitations to this study; the first one was that we discounted reviews that did not include a dominant topic. This reduced the population size and ignored reviews that might provide valuable insight. Second, both sentiment analysis and topic modeling inevitably have misclassification errors, especially when the user reviews can be misspelled or have a double meaning (sarcasm). The third limitation is that the topics from the topic modeling are influenced by words that carry sentiment; this causes some topics to carry sentimental meaning, such as *good staff*, rather than just the topic itself. Although this did not affect the accuracy of the results, it could cause the topics to be less useful.

Providing patient reviews in an easily inferable format through visualization tools to the public can foster competitiveness to improve among the NHS services. It is essential to provide analytical outcomes in an accessible format to general users to support patient autonomy and decision-making. The visualization tool developed in this study has the potential to rapidly and easily disseminate the findings of this study using maps and graph-based visual analytics. The visualization further enables users to view the feedback of an NHS service based on the key areas identified from the topics.

Conclusions

This study presents an automated system for the analysis of patient feedback based on NLP techniques and topic modeling.

The DSR methodology was adopted for this study to conduct the patient feedback analysis in 3 iterations. In the first iteration, sentiment analysis of the patient feedback corpus was performed followed by a topic modeling–based analysis of the corpus. In the third iteration, the sentiment scores and topics identified were further analyzed to associate the sentiment scores with the topics identified and categorize the corpus into good and bad corpora. Furthermore, we provided a data visualization interface with potential use for policy makers and associated stakeholders to review the performance of health care centers individually as well as by regions and identify the possible causes behind the performance of a health care center. The future work of our study aims to collect and analyze larger patient feedback databases to build more accurate analytical models. We are exploring artificial intelligence–based NLP models to include them in our analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Latent Dirichlet allocation–generated topics with their sentiment score.
[DOCX File , 18 KB - medinform_v10i4e29385_app1.docx]

Multimedia Appendix 2 [DOCX File , 15 KB - medinform_v10i4e29385_app2.docx]

Multimedia Appendix 3
Positive-sentiment topics (any review with a sentiment score >+0.2).
[DOCX File , 14 KB - medinform_v10i4e29385_app3.docx]

References

1. Defining patient experience. The Beryl Institute. URL: https://www.theberylinstitute.org/page/DefiningPatientExp [accessed 2022-03-16]
2. Manary MP, Boulding W, Staelin R, Glickman SW. The patient experience and health outcomes. N Engl J Med 2013 Jan 17;368(3):201-203. [doi: 10.1056/NEJMp1211775] [Medline: 23268647]
3. Coulter A. Leadership for patient engagement. Kings Fund. 2012. URL: https://www.kingsfund.org.uk/publications/leadership-engagement-for-improvement-nhs [accessed 2022-03-08]
4. Barello S, Graffigna G. Engaging patients to recover life projectuality: an Italian cross-disease framework. Qual Life Res 2015 May 6;24(5):1087-1096. [doi: 10.1007/s11136-014-0846-x] [Medline: 25373927]
5. Renedo A, Marston C. Healthcare professionals' representations of 'patient and public involvement' and creation of 'public participant' identities: Implications for the development of inclusive and bottom‐up community participation initiatives. J Community Appl Soc Psychol 2011 Apr 25;21(3):268-280 [FREE Full text] [doi: 10.1002/casp.1092]
6. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open 2013 Jan 03;3(1) [FREE Full text] [doi: 10.1136/bmjopen-2012-001570] [Medline: 23293244]
7. Dunderdale K. Patient experience improvement framework. National Health Services. 2018. URL: https://www.england.nhs.uk/publication/patient-experience-improvement-framework/ [accessed 2022-03-08]
8. Gluyas H. Patient-centred care: improving healthcare outcomes. Nurs Stand 2015 Sep 23;30(4):50-59. [doi: 10.7748/ns.30.4.50.e10186] [Medline: 26394978]
9. Davidson KW, Shaffer J, Ye S, Falzon L, Emeruwa IO, Sundquist K, et al. Interventions to improve hospital patient satisfaction with healthcare providers and systems: a systematic review. BMJ Qual Saf 2017 Jul 03;26(7):596-606 [FREE Full text] [doi: 10.1136/bmjqs-2015-004758] [Medline: 27488124]
10. Gleeson H, Calderon A, Swami V, Deighton J, Wolpert M, Edbrooke-Childs J. Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. BMJ Open 2016 Aug 16;6(8):e011907 [FREE Full text] [doi: 10.1136/bmjopen-2016-011907] [Medline: 27531733]
11. Coulter A, Fitzpatrick R, Cornwell J. The point of care. Measures of patients' experience in hospital: purpose, methods and uses. King's Fund. 2009. URL: https://www.researchgate.net/publication/230687403_The_Point_of_Care_Measures_of_Patients'_Experience_in_Hospital_Purpose_Methods_and_Uses [accessed 2022-03-08]
12. Beattie M, Murphy DJ, Atherton I, Lauder W. Instruments to measure patient experience of healthcare quality in hospitals: a systematic review. Syst Rev 2015 Jul 23;4(1):97 [FREE Full text] [doi: 10.1186/s13643-015-0089-0] [Medline: 26202326]
13. Managing patient feedback. National Health Services. URL: https://www.nhs.uk/about-us/managing-patient-feedback/ [accessed 2022-03-08]
14. Siegrist RB. Patient satisfaction: history, myths, and misperceptions. Virtual Mentor 2013 Nov 01;15(11):982-987 [FREE Full text] [doi: 10.1001/virtualmentor.2013.15.11.mhst1-1311] [Medline: 24257092]

15. Greaves F, Pape UJ, King D, Darzi A, Majeed A, Wachter RM, et al. Associations between web-based patient ratings and objective measures of hospital quality. Arch Intern Med 2012 Mar 12;172(5):435-436. [doi: 10.1001/archinternmed.2011.1675] [Medline: 22331980]
16. Bahja M, Lycett M. Identifying patient experience from online resources via sentiment analysis and topic modelling. In: Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. 2016 Presented at: UCC '16: 9th International Conference on Utility and Cloud Computing; December 6 - 9, 2016; Shanghai China p. 94-99. [doi: 10.1145/3006299.3006335]
17. Bahja M, Razaak M. Automated analysis of patient experience text mining using a design science research (DSR) approach. In: Proceedings of the BUSTECH 2018: The Eighth International Conference on Business Intelligence and Technology. 2018 Presented at: BUSTECH 2018: The Eighth International Conference on Business Intelligence and Technology; February 18 - 22, 2018; Barcelona, Spain URL: https://thinkmind.org/index.php?view=article&articleid=bustech_2018_2_10_98002
18. Peffers K, Tuunanen T, Gengler C, Rossi M. The design science research process: a model for producing and presenting information systems research. ResearchGate. 2006. URL: https://www.researchgate.net/publication/238077290_The_design_science_research_process_a_model_for_producing_and_presenting_information_systems_research [accessed 2022-03-08]
19. Gregor S, Hevner AR. Positioning and presenting design science research for maximum impact. MIS Q 2013 Feb 2;37(2):337-355. [doi: 10.25300/misq/2013/37.2.01]
20. Ratings and reviews API. NHS APIs. URL: https://developer.api.nhs.uk/nhs-api/documentation/comments [accessed 2022-03-08]
21. NHS website syndicated content: standard licence terms. National Health Services. URL: https://apimgmtst3acoupair9misya.blob.core.windows.net/content/MediaLibrary/Terms/NHS-website-syndication-terms.pdf [accessed 2022-08-03]
22. Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013 Apr;56(4):82-89. [doi: 10.1145/2436256.2436274]
23. Feinerer I. A text mining framework in R and its applications. ResearchGate. 2008. URL: https://www.researchgate.net/publication/277070985_A_text_mining_framework_in_R_and_its_applications [accessed 2022-03-08]
24. Advanced usage: overriding models and the Blobber Class. TextBlob. URL: https://textblob.readthedocs.io/en/dev/advanced_usage.html#sentiment-analyzers [accessed 2022-03-08]
25. D'Andrea A, Ferri F, Grifoni P, Guzzo T. Approaches, tools and applications for sentiment analysis implementation. Int J Comput Appl 2015 Sep 17;125(3):26-33. [doi: 10.5120/ijca2015905866]
26. Su Q, Xiang K, Wang H, Sun B, Yu S. Using pointwise mutual information to identify implicit features in customer reviews. In: Proceedings of the International Conference on Computer Processing of Oriental Languages. 2006 Presented at: International Conference on Computer Processing of Oriental Languages; December 17-19, 2006; Singapore. [doi: 10.1007/11940098_3]
27. Boushaki SI, Kamel N, Bendjeghaba O. High-dimensional text datasets clustering algorithm based on Cuckoo search and latent semantic indexing. J Info Know Manag 2018 Oct 02;17(03):1850033. [doi: 10.1142/s0219649218500338]
28. Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat 2007 Jun 1;1(1):17-35. [doi: 10.1214/07-aoas114]
29. He J, Hu Z, Berg-Kirkpatrick T, Huang Y, Xing EP. Efficient correlated topic modeling with topic embedding. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 Presented at: KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13 - 17, 2017; Halifax NS Canada p. 225-233. [doi: 10.1145/3097983.3098074]
30. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2018 Nov 28;78(11):15169-15211. [doi: 10.1007/s11042-018-6894-4]
31. Arun R, Suresh V, Madhavan C, Murthy M. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2010 Presented at: Pacific-Asia Conference on Knowledge Discovery and Data Mining; June 21-24, 2010; Hyderabad, India. [doi: 10.1007/978-3-642-13657-3_43]
32. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text]
33. Topic modeling. Machine Learning for Language Toolkit. URL: http://mallet.cs.umass.edu/topics.php [accessed 2020-12-01]
34. Phrase (collocation) detection. Gensim. URL: https://radimrehurek.com/gensim/models/phrases.html [accessed 2019-09-06]
35. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015 Presented at: WSDM 2015: Eighth ACM International Conference on Web Search and Data Mining; February 2 - 6, 2015; Shanghai China p. 399-408. [doi: 10.1145/2684822.2685324]
36. Overview of Azure Cloud Services (classic). Microsoft. 2021. URL: https://docs.microsoft.com/en-us/azure/cloud-services/cloud-services-choose-me [accessed 2020-01-07]
37. Node.js. URL: https://nodejs.org/en/ [accessed 2022-03-08]
38. Vue.js: The Progressive JavaScript Framework. URL: https://vuejs.org/ [accessed 2022-03-08]
39. Google Maps Platform. Google. URL: https://developers.google.com/maps [accessed 2020-07-01]

40. Juszczak J. vue-chartjs. URL: https://vue-chartjs.org/ [accessed 2022-03-08]

Abbreviations

API: application programming interface
CTM: correlated topic modeling
DSR: design science research
LDA: latent Dirichlet allocation
LSI: latent semantic indexing
NB: naïve Bayes
NHS: National Health Service
NLP: natural language processing

Edited by C Lovis; submitted 05.04.21; peer-reviewed by R Hammad, YT Choi; comments to author 24.04.21; revised version received 08.09.21; accepted 04.12.21; published 11.04.22.

Please cite as:
Alexander G, Bahja M, Butt GF
Automating Large-scale Health Care Service Feedback Analysis: Sentiment Analysis and Topic Modeling Study
JMIR Med Inform 2022;10(4):e29385
URL: https://medinform.jmir.org/2022/4/e29385
doi:10.2196/29385
PMID:35404254

Original Paper

Generation of a Fast Healthcare Interoperability Resources (FHIR)-based Ontology for Federated Feasibility Queries in the Context of COVID-19: Feasibility Study

Lorenz Rosenau[1], MSc; Raphael W Majeed[2], MSc; Josef Ingenerf[1], PhD; Alexander Kiel[3,4], BSc; Björn Kroll[1], PhD; Thomas Köhler[4,5]; Hans-Ulrich Prokosch[6], PhD; Julian Gruendner[6], MA, MSc

[1]IT Center for Clinical Research, Lübeck, Germany

[2]Institute for Medical Informatics, University Clinic Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

[3]Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig, Germany

[4]Federated Information Systems, German Cancer Research Center, Heidelberg, Germany

[5]Complex Data Processing in Medical Informatics, Medical Faculty Mannheim, Mannheim, Germany

[6]Chair of Medical Informatics, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Lorenz Rosenau, MSc
IT Center for Clinical Research
Gebäude 64, 2.OG, Raum 05
Ratzeburger Allee 160
Lübeck, 23562
Germany
Phone: 49 451 3101 5636
Email: lorenz.rosenau@uni-luebeck.de

Abstract

Background: The COVID-19 pandemic highlighted the importance of making research data from all German hospitals available to scientists to respond to current and future pandemics promptly. The heterogeneous data originating from proprietary systems at hospitals' sites must be harmonized and accessible. The German Corona Consensus Dataset (GECCO) specifies how data for COVID-19 patients will be standardized in Fast Healthcare Interoperability Resources (FHIR) profiles across German hospitals. However, given the complexity of the FHIR standard, the data harmonization is not sufficient to make the data accessible. A simplified visual representation is needed to reduce the technical burden, while allowing feasibility queries.

Objective: This study investigates how a search ontology can be automatically generated using FHIR profiles and a terminology server. Furthermore, it describes how this ontology can be used in a user interface (UI) and how a mapping and a terminology tree created together with the ontology can translate user input into FHIR queries.

Methods: We used the FHIR profiles from the GECCO data set combined with a terminology server to generate an ontology and the required mapping files for the translation. We analyzed the profiles and identified search criteria for the visual representation. In this process, we reduced the complex profiles to code value pairs for improved usability. We enriched our ontology with the necessary information to display it in a UI. We also developed an intermediate query language to transform the queries from the UI to federated FHIR requests. Separation of concerns resulted in discrepancies between the criteria used in the intermediate query format and the target query language. Therefore, a mapping was created to reintroduce all information relevant for creating the query in its target language. Further, we generated a tree representation of the ontology hierarchy, which allows resolving child concepts in the process.

Results: In the scope of this project, 82 (99%) of 83 elements defined in the GECCO profile were successfully implemented. We verified our solution based on an independently developed test patient. A discrepancy between the test data and the criteria was found in 6 cases due to different versions used to generate the test data and the UI profiles, the support for specific code systems, and the evaluation of postcoordinated Systematized Nomenclature of Medicine (SNOMED) codes. Our results highlight the need for governance mechanisms for version changes, concept mapping between values from different code systems encoding the same concept, and support for different unit dimensions.

Conclusions: We developed an automatic process to generate ontology and mapping files for FHIR-formatted data. Our tests found that this process works for most of our chosen FHIR profile criteria. The process established here works directly with FHIR profiles and a terminology server, making it extendable to other FHIR profiles and demonstrating that automatic ontology generation on FHIR profiles is feasible.

(JMIR Med Inform 2022;10(4):e35789) doi:10.2196/35789

KEYWORDS

federated queries; feasibility study; Fast Healthcare Interoperability Resource; FHIR Search; CQL; ontology; terminology server; query; feasibility; FHIR; terminology; development; COVID-19; automation; user interface; map; input; hospital; data; Germany; accessibility; harmonized

Introduction

Background

Researchers require data to test, refine, and improve their models. Historically in health care, these data have often only been accessible and discoverable locally. Due to different protocols, proprietary solutions, and missing terminology, there is a lack of standardization to promote interoperability and data reuse [1].

In a national effort, the Medical Informatics Initiative (MII) in 2017 started to establish a national research platform for health care professions [2]. Local data integration centers (DICs) collect the vast amount of health care data from the clinics and make them accessible across institutional boundaries. The DICs provide different services, such as data integration, data harmonization, standardized data repositories, consent management, and ID management, and form the backbone of a cross-institutional research network.

Data harmonization is achieved by applying Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR), which is an interoperability standard for health care information [3]. It defines a common health care business entity model with *Resources* as basic building blocks. Each *Resource* has a defined set of data elements, constraints, and relationships to other *Resources*. Common *Resources* relevant to clinical researchers are *Patient*, *Observation*, *Condition*, *Procedure*, *MedicationStatement*, *Consent*, and *Immunization*. FHIR profiles can further constrain and extend the predefined *Resources* for specific use cases.

The COVID-19 pandemic revealed the urgency of addressing the interoperability challenge [4]. The German Corona Consensus Dataset (GECCO) [5] and its representation in FHIR profiles were developed to address the semantic interoperability challenge on a national level.

GECCO consists of 83 data elements defined in FHIR profiles that characterize COVID-19 patients according to their medical history, findings, demographics, laboratory values, medications, symptoms, therapy, and vital signs. Each profile's *Bindings* to *ValueSets* (defined sets of medical terminology) that reference the *CodeSystems* Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), *International Classification of Diseases and Related Health Problems, 10th edition*, German version (ICD-10-GM), and Anatomical Therapeutic Chemical (ATC) [5] define the medical terms associated with COVID-19 patients within the German health care system. The data set is under ongoing development.

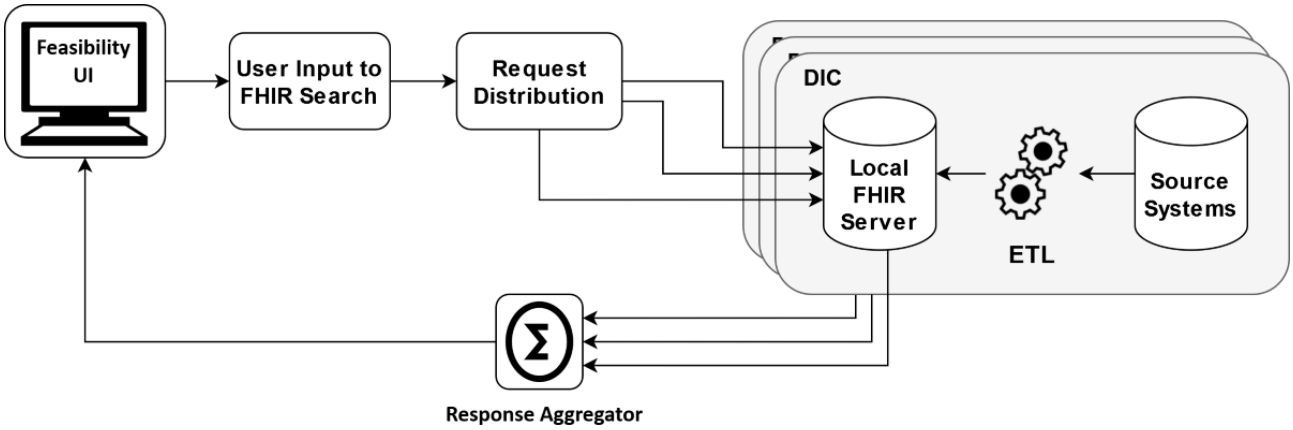
In the CODEX project funded by the German Federal Ministry of Education and Research (BMBF), the existing infrastructural progress of the MII is the foundation to create a web-based federated query tool, which researchers can use for cohort discovery/feasibility queries based on the GECCO data model.

Within the CODEX feasibility architecture (Figure 1), all German university hospitals extract, transform, and load (ETL) their COVID-19 patient data from their primary source systems to a local FHIR server in GECCO format. Feasibility queries created in the central CODEX feasibility user interface (UI) are forwarded via the CODEX feasibility platform to the decentralized FHIR server within the DICs. Their responses are then transported back to the feasibility UI and displayed to the user, anonymized and aggregated. The detailed architecture is described in a separate publication [6].

The feasibility platform developed within the CODEX project is independent of the COVID-19 use case. Within the FHIR server, arbitrary data can be stored if ETL processes exist to convert the clinical source systems data to FHIR. Furthermore, the query languages (FHIR Search and Clinical Quality Language [CQL]) used at the DICs are universally applicable for arbitrary FHIR data. The highly reusable nature of the infrastructure lends itself well to developing a UI that is use-case-independent. The structure of feasibility queries is consistent—only the use-case-specific query criteria need to be identified. Therefore, for our use case, the data elements within GECCO need to be provided to the user as query criteria.

Extracting criteria from structured data based on a clinical data model for a visual representation on a query interface was also performed by Haarbrandt et al [7] for the openEHR format (where "EHR" refers to "electronic health record"). Contrary to their approach, we keep the FHIR data in their existing format and do not rely on ETL processes. Similar to other federated approaches [8-10], we create feasibility queries centrally and distribute them to the clinical sites. In contrast to them, our feasibility platform is based on FHIR profiles.

Figure 1. CODEX feasibility architecture. DIC: data integration center; ETL: extract, transform, and load; FHIR: Fast Healthcare Interoperability Resources; UI: user interface.



Aim

The complex nature of FHIR profiles makes them unsuitable as a direct interaction format for researchers. This study investigates the use of FHIR profiles, using the GECCO profile as an example, to automatically generate an ontology that provides a generic UI with all the information needed to create feasibility queries and execute them at the hospital sites. We use the term "ontology" following Informatics for Integrating Biology & the Bedside (i2b2; i2b2 Foundation Inc) to refer to hierarchically structured concepts that allow users to create queries using the concepts as criteria [7].

Methods

Overview

The aim of generating an ontology is to make criteria findable and identifiable by researchers. These criteria are often independent of how data are stored and processed on a technical level. To bridge this gap, this study investigated not only how to generate an ontology for a UI but also how a mapping and a terminology tree file can be automatically generated to support FHIR request generation.

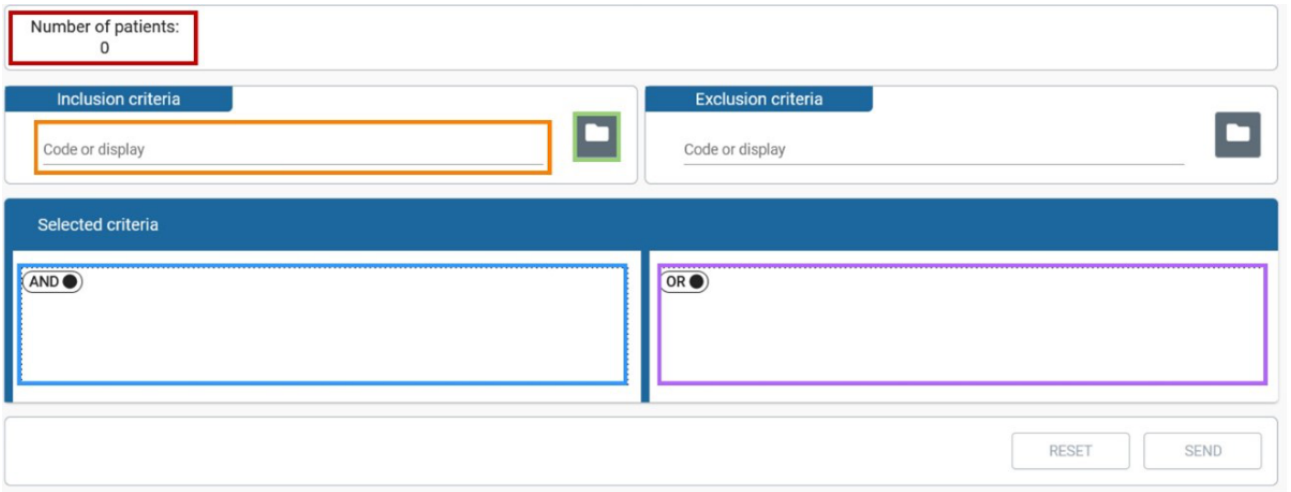
Thus, we divided the investigation of the problem into 2 parts:

- Creating UI profiles for the visual representation in the UI
- Creating a mapping and a terminology tree for query translation

UI Profiles

The UI (Figure 2), designed for feasibility queries, allows the user to select inclusion and exclusion criteria. The criteria can be chosen from a tree representation (green) or searched for directly (orange). Inclusion and exclusion criteria are presented in a drag-and-drop area where different criteria can be joined using the Boolean AND and OR operations and moved from inclusion to exclusion and vice versa (blue and purple). The represented concepts can be stand-alone or further specified by the user with a value while defining the query. The criteria the user can choose from in the CODEX project are based on GECCO.

Figure 2. The Codex feasibility UI containing widgets to choose criteria and to create suitable queries. UI: user interface.



The GECCO profiles are defined as FHIR *StructureDefinitions* and can be obtained from Simplifier [11]. Each profile can be regarded as a blueprint of possible *Resource* instance data stored in the DICs.

A profile analysis must provide uniquely identifiable elements and values of interest that define the criteria to create the ontology for the user. Manual maintenance of such an ontology would be a time-consuming, error-prone, and laborious task

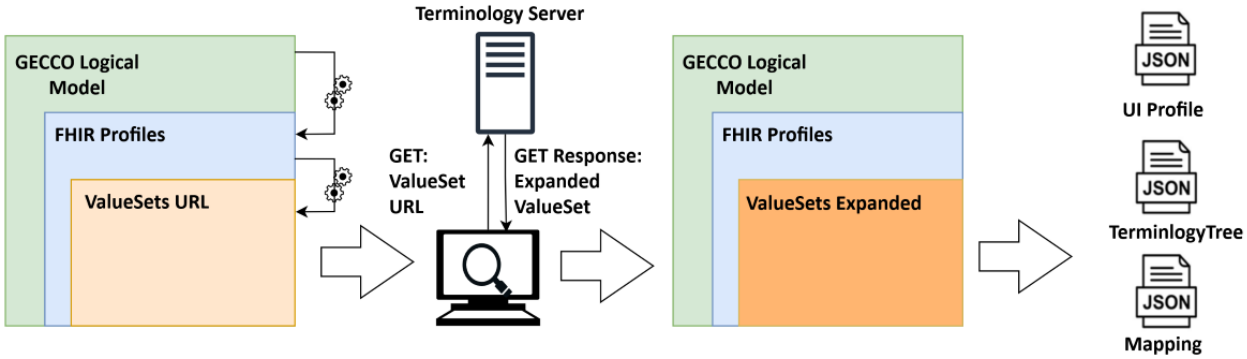
[12]. Given the structured nature of the FHIR profiles, an automated approach can be used to generate the ontology. For this purpose, we implemented a Python script [13], which creates a JavaScript Object Notation (JSON) representation of the ontology—the UI profiles (see Figure 3 for an excerpt). This representation puts all criteria in a hierarchical context using a *children* element for each criterion and provides the UI with all the necessary information to display each criterion. If the *children* element is empty, the criterion is a *leaf* criterion, which does not need to be expanded further.

Figure 4 illustrates the entire program's procedure. Besides the UI profile, a mapping and a terminology tree were created.

Figure 3. UI profile excerpt. UI: user interface.


```
{
    "display": "Anamnese / Risikofaktoren",
    "id": "bc8c8790-85cd-4b53-bf7e-3b4fc162fele",
    "leaf": false,
    "selectable": false,
    "termCode": {
        "code": "Anamnese /Risikofaktoren",
        "display": "Anamnese / Risikofaktoren",
        "system": "num.codex"
    },
    "timeRestrictionAllowed": false,
    "valueDefinitions": []
    "children": [
        {
            "children": [   "..."   ],
            "display": "Aktive Tumor-/Krebserkrankungen",
            "id": "6c9f241a-1c39-4f26-8cc2-c2574ca3183e",
            "leaf": false,
            "selectable": true,
            "termCode": {
                "code": "363346000",
                "display": "Malignant neoplastic disease (disorder)",
                "system": "http://snomed.info/sct"
            },
            "timeRestrictionAllowed": false,
            "valueDefinitions": []
        },
    "..."
}
```


Figure 4. Processing of the GECCO profile to UI profile, mapping, and terminology tree. First, the FHIR profiles are identified within the *LogicalModel*. Next, the *ValueSets* defined in the *Bindings* of specific attributes within the FHIR profiles are identified. Afterward, the *ValueSets* are expanded utilizing a terminology server. Finally, the combined information from the *LogicalModel*, the FHIR profiles, and the expanded *ValueSets* gets processed and converted to the UI profile, the mapping, and the terminology tree. FHIR: Fast Healthcare Interoperability Resources; GECCO: German Corona Consensus Dataset; JSON: JavaScript Object Notation; UI: user interface.



In addition to the *StructureDefinitions*, the GECCO profile provides a *LogicalModel*. FHIR logical models serve the purpose of collecting requirements from medical experts without having to adhere to the FHIR specifications in the early stages of profile development. In later stages, the elements within the *LogicalModel* can be mapped to the *StructureDefinitions*. For us, the JSON representation of the *LogicalModel* served to identify the categories of our UI. For each category, the *LogicalModel* further defined a set of logical criteria. The name of each criterion was then used to identify the respective profile representing the criteria. Not every GECCO profile needs to be handled individually. The implementation effort can be drastically reduced by grouping all profiles based on the FHIR *ResourceType*.

Each criterion is specified by a code from a terminology system. An optional value allows further restricting the criteria. If no value is provided, the existence of the code is the criterion.

An in-depth analysis of the FHIR profiles allowed us to identify the attributes that specify the criteria and their values. Table 1 displays the attributes of a FHIR profile, which specify the criteria and the values for each FHIR *ResourceType*. In total, 75 (90%) of the 83 defined profiles could be represented in this fashion.

Table 1. Identified attributes that specify the concepts and values for the criteria.

ResourceType	Criteria-specifying attribute	Value-specifying attribute	Example
Condition	code		Type 2 diabetes mellitus
Observation (concept)	code	value (*CodeableConcept*)	Sex assigned at birth: female
Observation (quantity)	code	value (*Quantity*)	Weight: 70 kg
Procedure	code		Plain radiography (procedure)
MedicationStatement	code		Product containing antipyretic
Immunization	vaccineCode		Typhus vaccine (product)
DiagnosticReport	code	conclusion	Diagnostic imaging study: radiological finding characteristic for COVID-19
Specimen	type		Blood specimen

Some profiles with the same *ResourceType* do not hold the information on the value in the same attribute. For these cases, additional heuristics or corner cases need to be established. One reoccurring case in this regard is the representation of the *ObservationResource*. FHIR does not differ between *Observations* that have recorded a concept or a value. For example, the concept of smoking status is defined as an *Observation* with values indicating the smoking frequency. The body height is also defined as an *Observation* but has a quantity as a value. Therefore, different UI profiles and mappings are needed.

The profile itself is not a criterion. Instead, the profile's criteria-specifying attribute (see Table 1) defines the set of criteria, and the profile specifies how all criteria within this set will be modeled.

After identifying the *ResourceType*, the set of criteria and possible values for each criterion can be resolved using the *Bindings* of each specifying attribute. Each *Binding* contains the canonical URL of a *ValueSet*. A *ValueSet* defines a set of medical terms from medical terminology, such as ICD-10. An instance of the Ontoserver, a terminology server [14] based on the FHIR standard, administers all *ValueSets* from the GECCO profile. After identifying the *ValueSet*, the available values can be obtained from the terminology server using the *expand* operation. Each concept in the *ValueSet* has a unique combination of code and system, which identifies the criterion. The concepts are represented in a list. To build our ontology, we derived the hierarchy of codes based on the is-a relationship between them. We further enriched our ontology with information about how the criteria should be represented (ie, which criterion is selectable).

To illustrate the process, take the criteria group “Chronic Lung Disease” with the parent category “Anamnesia/Risk Factors.” The profiles JSON is analyzed and based on the field *ResourceType*, identified as a *Condition* whose attribute “code” (contains a code and a system) defines the criterion. Other attributes are in this case not of primary interest to the researcher and can be ignored during query processing or set to specific values for the most common research interest, like only searching for verified conditions.

Valid codes can be obtained from SNOMED-CT and ICD-10-GM *ValueSets*. Currently, valid codes are only displayed for codes from a single *CodeSystem* due to the potential confusion caused by the overlap of concepts between *CodeSystem* (ie, sleep apnea is part of ICD-10 and SNOMED CT). The ICD-10-GM *CodeSystem* is chosen because of its broader adaptation in clinics. The *ValueSet* is transformed into a tree structure based on the subsumption relations within the terminology and appended below the “Chronic Lung Disease” node.

FHIR Search/CQL

Between its visual representation and the execution as a FHIR Search request at the university hospital sites, the feasibility query created in the UI is sent to the backend in an intermediate data format. The intermediate query format was developed within the CODEX project and is named Structured Query (SQ).

Like the UI, the SQ is composed of 2 parts:

- The inclusion criteria are in conjunctive normal form without negation.
- The exclusion criteria are in disjunctive normal form without negation.

They are combined in an AND NOT expression:

SQ=inclusion criteria (CNF) ¬ exclusion criteria (DNF)

The use of an intermediate format simplifies the translation into multiple query languages. FHIR *Resources* can be requested using FHIR Search or CQL. FHIR Search uses GET requests to obtain *Resources* from an FHIR server. All *Resources* define a set of search parameters that can be used to filter the search result.

FHIR Search has limitations in its expressiveness. It requires defined search parameters and cannot express inclusion and exclusion criteria in a single query [15].

Although these issues have been overcome within the CODEX project through workarounds including custom search parameters, multiple FHIR Search requests, and combination logic of the results, CQL presents a promising solution to overcome the limitations of FHIR Search [16].

Mapping

To allow for a high degree of modularity we applied the software design pattern Separation of Concerns [17]. This allowed for independent development of the components and provided more flexibility to adjust to individual sites' existing infrastructure and future developments. The UI is separated from the query process and the query language, allowing high maintainability. Therefore, the UI profiles do not hold information on the underlying FHIR data model or the query languages. Furthermore, the hierarchic information is not transferred in the SQ, allowing for independent ontology development.

Therefore, the lost information about the *Resources* and their search parameters needed to create the FHIR Search request at the clinical server side must be reintroduced.

To achieve this, we created a mapping for each criterion (Figure 5), storing all information needed to translate the SQ into FHIR Search and CQL requests.

Figure 5. Mapping entry for "Chronic Lung Disease." The search parameter for the code identifying the criterion is "code." The value of *verificationStatus* is fixed to "confirmed."


```
{
    "fhirResourceType": "Condition",
    "fixedCriteria": [
        {
            "fhirPath": "verificationStatus",
            "searchParameter": "verification-status",
            "type": "coding",
            "value": [
                {
                    "code": "confirmed",
                    "display": "confirmed",
                    "system": "http://terminology.hl7.org/CodeSystem/condition-ver-status"
                }
            ]
        }
    ],
    "termCode": {
        "code": "413839001",
        "display": "Chronic lung disease",
        "system": "http://snomed.info/sct"
    },
    "termCodeSearchParameter": "code"
}
```


Again, we used the same process as established previously to generate the UI profiles. Instead of rendering the codes for the criteria- and value-specifying attributes, we linked the codes and the search parameter and FHIR paths for the same criteria.

Utilizing the criteria code as a key, we specified the search parameters for the code and the value.

Not all attributes of a FHIR profile have a default search parameter, especially all *Extensions* as they are not part of the official FHIR standard. To handle these cases, additional (custom) search parameters needed to be defined, added to the FHIR server, and referred to in our mapping.

We further defined so-called fixed criteria to restrict attributes not available to the user by setting their search parameter to a predefined value. This is necessary, for example, to only search for confirmed diagnoses.

For the chronic lung disease example, a mapping entry was created for each chronic lung disease code with the corresponding information that the code can be found within the resource *Condition* under the search parameter "code" with a fixed criterion "verification-status" with the value "confirmed" (see Figure 5).

Criteria that are not *leaves* in the ontology tree represent all criteria that descend from it. The subcriteria are not sent in the SQ and need to be resolved at the clinical sites. Due to the lack of terminology servers, we provided the terminology tree JSON file, which represents the UI profiles reduced to only hierarchic

information between codes. A terminology tree consists of nodes with 2 properties: the code that identifies the concept within the tree and a list of child nodes.

Results

Corner Cases

The established process can parse all profiles defined in GECCO. However, in an in-depth analysis of the GECCO profiles, we identified 7 corner cases needing explicit handling, increasing the implementation effort. Table 2 lists the issues preventing the handling based on *ResourceType*. Explicit handlings were implemented for each case.

Using the explicit and *ResourceType*-based mapping, we successfully created the UI profiles, mapping, and terminology tree for the additional 7 corner cases, thus covering a total of 82 (99%) of 83 profiles. Only the date of birth was excluded due to privacy concerns but could have been implemented in a similar manner.

Examples of the feasibility UI with the loaded UI profiles and an example query can be found in Multimedia Appendix 1.

The overall architecture utilized the results as shown in Figure 6.

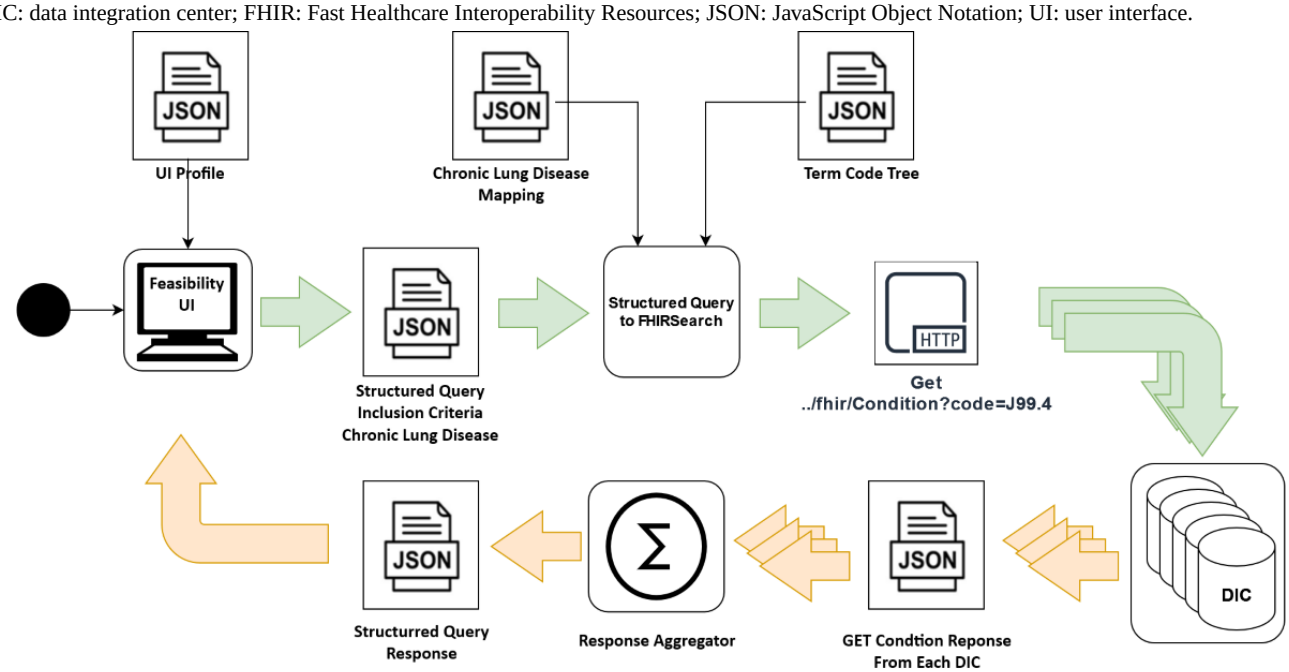
The criteria were selected and combined into a feasibility query based on the UI profiles. The resulting SQ was sent to a back-end component, which translated the SQ utilizing the mapping. The resulting FHIR Search requests were distributed to all DICs at the clinical sites and executed on the GECCO-harmonized data using the mapping and terminology tree we generated. The responses were aggregated, anonymized, and sent back to the feasibility UI to display the result.

Table 2. Corner cases, by their profile name and *ResourceType*, and the issue preventing the default handling.

Profile	*ResourceType*	Issue preventing default handling
Sequential Organ Failure Assessment (SOFA)	*Observation*	The value of the SOFA score is stored in value[integer], not in value[quantity].
History of Travel	*Observation*	The information of interest is stored in a component.
Systolic/Diastolic Blood Pressure	*Observation*	The information of interest is stored in a component. Contrary to the "History of Travel," "Systolic/Diastolic Blood Pressure" is stored as a quantity, not as a concept.
Covid-19 Symptoms	*Condition*	For COVID-19 symptoms, we decided that the severity should also be settable by the researcher as a value.
Ethnic Group	*Extension*	The ethnic group is an extension and needs a specific search parameter and FHIR[a] path.
Age	*Extension*	Age is an extension and needs a specific search parameter and FHIR path.

[a]FHIR: Fast Healthcare Interoperability Resources.

Figure 6. Activity diagram showcasing the creation and execution of the feasibility request based on the UI profiles in the CODEX feasibility architecture. DIC: data integration center; FHIR: Fast Healthcare Interoperability Resources; JSON: JavaScript Object Notation; UI: user interface.



Evaluation

At the time of writing this publication, the DICs were still under development, and the ETL processes to fill the FHIR servers with real-world GECCO data have yet to be rolled out. Many hospital sites use the electronic data capture tool REDCap [18] to collect COVID-19 patient data and the ODM2FHIR tool [19] to transform the data to FHIR. For our automated and manual tests, we used this toolchain to create our test patients. The manual tests were conducted by selecting logical combinations from the criteria defining the test patient in the UI. In addition, we generated SQs that request each criterion and should return our test patient as a result. The test data and the generated SQs are available in Ref. [20].

In 6 (7%) of 84 conducted manual tests, a discrepancy between the test data encoding and the available elements in the UI made it impossible to obtain the data of interest. The 6 discrepancies were caused by 4 different sources of errors:

- SNOMED CT postcoordination: SNOMED CT makes it possible to specify concepts (eg, defining the body side of a finding) using postcoordinated expression (PCE) [21]. PCE-coded concepts represent a subset of a non-PCE-coded concept but are not part of expanded value sets if not explicitly defined. In consequence, only the non-PCE-coded concept is available in the UI.
- GECCO version discrepancies: Although GECCO version 1.0.4 was used as the basis for the UI implementation, the test data is still based on the previous version 1.0.3. This discrepancy sometimes results in different coding for the concepts.
- Unit definitions: The *LogicalModel* of GECCO defines units for all quantitative values. The current implementation does not allow converting between units. Users must search the unit according to the test data, leading to errors in 2 cases where the unit is unavailable.
- *CodeSystem* discrepancies: Although the GECCO profile allows for values from different *CodeSystem*, we reduced this complexity to values from a single *CodeSystem*. Not for every value does a corresponding code in all *CodeSystem* exist. Consequently, some codes in the test data are not available in the UI.

Discussion

Principal Findings

We presented the automatic generation of an ontology for a federated feasibility search tool and the necessary information to translate an intermediate query format to FHIR Search and CQL. We based the generation of the ontology, and the mapping, on FHIR profiles, allowing us to generalize our method to FHIR profiles, which represent a concept with a unique identifying code and an optional value. We successfully implemented UI profiles (UI representations) as well as the mapping for all criteria from GECCO and verified our solution based on an independently developed test patient.

We use FHIR data in their original format while simultaneously representing the concepts as criteria in a simplified model for the end user, resulting in a reduced technical burden, which improves usability. Other ETL processes on the FHIR data are unnecessary. Further, we generated the ontology automatically and did not rely on manual maintenance. Consequently, the development time of an ontology can be drastically reduced, and the ontology can be adapted rapidly to version changes of the data set.

Related Work

The development of a feasibility portal for medical health data poses an ill-structured problem. A wide opportunity space holds solutions in different architectures, data formats, query languages, and tooling.

A federated approach is the greatest common feature between existing feasibility solutions to overcome legal boundaries and ensure privacy protection on sensitive health care data. For proprietary, i2b2, and Observational Medical Outcomes Partnership (OMOP) data, solutions exist that provide researchers with an ontology-based UI [9,10,22]. These platforms can also be utilized for FHIR and openEHR data but require additional ETL processes [7,23]. The Leaf project [8] presents an alternative approach by using a model agnostic query system for medical data stored in Structured Query Language (SQL) databases. Like our approach, an ontology holds the information on the criteria available to the user, and similar criteria are mapped to WHERE clauses for SQL statements. To apply their query system to FHIR requires a flat representation of the FHIR *Resources* in a SQL database. As the used FHIR servers at the DICs do not store flattened representations of the FHIR profiles and an additional representation in flattened form would cause data redundancy, their solution could not be applied to our problem. Regardless, an ontology and a mapping would have also been needed to utilize the Leaf approach. Other existing solutions utilizing the FHIR standard for federated feasibility queries rely on computer scientists to transfer their research questions to FHIR Search, CQL, or SQL [24,25]. Existing FHIR-based federated feasibility query tools with a graphical UI, developed for health care professionals, rely on manual curation of search criteria [26,27]. Manual curation is a laborious task and can take years.

With the presented work, we provide a solution for creating an ontology based on FHIR profiles suitable for medical professionals to create and execute federated feasibility queries for data in FHIR format.

Lessons Learned

The presented methodology relies on the extensive investigation of the FHIR profiles. Often, the expertise in those lies with the domain experts and modelers. Software developers must not only identify handling for individual *Resources* based on FHIR types but also discover all corner cases. A more interdisciplinary team could facilitate and shorten the development process. The presented implementation for GECCO can act as a starting point for other FHIR profiles. Developers need to add handling for *ResourceTypes* that are not yet implemented and add corner cases for profiles that do not align with the default handling.

The development and especially the delivery of the ontology rely on the infrastructure at the clinical sites. The Blaze FHIR server [28] implementation utilized in this project allowed the

usage of CQL and custom search parameters. In contrast, a lack of terminology servers at the sites resulted in the need to make the ontology available in a proprietary format and prevented using the *below* modifier a terminology server offers. In the future, the definition of custom search parameters should be part of the profiling process to ensure that the criteria defined in GECCO are queryable.

Limitations

Further improvements can be made to our solution to address the issues found. The SNOMED CT postcoordination limitations can be addressed by using the *below* modifier in FHIR Search requests. The *below* modifier resolves the is-relation between the PCE and the non-PCE equivalent but requires a SNOMED CT *CodeSystem* at every site.

Given the ongoing development and fixes in GECCO, our static approach for the UI profiles currently limits the use to a single version. Given the federated nature of the project, we cannot guarantee that every site uses the newest version. Therefore, support of multiple versions would be helpful. Improvements can be made by utilizing the terminology server in conjunction with versioning at run time to create the UI profiles semidynamically.

For usability, the units provided should be converted to the units used at each site during query execution. Research efforts to address this issue can be found in Ref. [29].

The flexible use of values from different *CodeSystems* represents the most significant challenge, as it cannot be solved on a purely technical level. Reducing the values provided to values from a single *CodeSystem* serves to simplify the presentation for the user. Concepts repeated in different *CodeSystems* are listed only once in the UI (eg, sleep apnea is available in ICD-10-GM and SNOMED CT but can only be selected as an ICD-10-GM concept). A mapping between all codes would be necessary to support both code systems. This mapping requires medical expertise as not all concepts can be as directly matched as the example. Stricter profiling with values limited to a single *CodeSystem* would have resulted in a higher workload at each site but improved organizational interoperability. Narrowing the optionality reduces the complexity, ultimately leading to better interoperability [30].

Future Directions

The high adaptability of the developed platform and the presented methodology open possibilities for a wide range of future work. Applying the presented approach to other FHIR data sets is part of ongoing work in the successor project of CODEX, ABIDE [31], where the same approach is applied to the MII core data set [32]. For cancer research, the presented approach could also be applied to the data model in Ref. [33].

Regarding FHIR, we want to expand the code value representation by establishing attribute filters that further refine the criteria using multiple FHIR *Resource* attributes.

Beyond FHIR, it would also be of interest to test the adaptability of our approach to other structured health care data. Primarily dependent on the mapping capabilities, we see the potential to use the SQ as an intermediate query language for FHIR and other query languages (ie, Archetype Query Language [AQL]) [34]. Previous research work [35] indicates the feasibility of this idea.

The current representation of the ontology is a proprietary format developed within this project. For better exchange, it should be investigated whether the features of a terminology server can be used to exchange the developed ontology in the standardized FHIR format (ie, using a structure map for the mapping) and dynamically load it from there.

Finally, a mapping between complex FHIR *Resources* and simplified interface patterns should be further investigated. The Release 5 draft of the FHIR standard introduces interface patterns, which could abstract a simplified representation from the FHIR *Resource*. Combined with the FHIR mapping language, a simpler resource data model for querying could be developed by domain experts rather than software developers.

Conclusion

We demonstrated an automated process to generate an ontology for feasibility criteria based on GECCO profiles, showcasing the feasibility of our approach for FHIR-profiled data. We described how to obtain user-relevant data from the FHIR profiles and how to use the same information to create a mapping to translate an intermediate query language to CQL and FHIR Search.

The underlying platform has been deployed across 33 university hospitals in Germany. Test data were used to evaluate our approach and demonstrate its validity.

We see great generalization potential not only for other FHIR profiles but also for structured health care data in general.

Acknowledgments

This work was performed in fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (JG).

The project was funded by the German Federal Ministry of Education and Research (BMBF; Grant 01KX2021).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Example of loaded UI profiles in the generic feasibility UI and an example query. UI: user interface.
[PDF File (Adobe PDF File), 406 KB - medinform_v10i4e35789_app1.pdf]

References

1. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019 Aug 20;2(1):79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
2. Semler S, Wissing F, Heyder R. German Medical Informatics Initiative: a national approach to integrating health data from patient care and medical research. Methods Inf Med 2018 Jul 17;57(S 01):e50-e56. [doi: 10.3414/me18-03-0003]
3. Summary - FHIR v4.0.1. URL: http://hl7.org/fhir/summary.html [accessed 2021-09-13]
4. Weber S, Heitmann KU. Interoperability in healthcare: also prescribed for digital health applications (DiGA). Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021 Oct 16;64(10):1262-1268 [FREE Full text] [doi: 10.1007/s00103-021-03414-w] [Medline: 34532746]
5. Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 2020 Dec 21;20(1):341 [FREE Full text] [doi: 10.1186/s12911-020-01374-w] [Medline: 33349259]
6. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Rosenau L, et al. Architecture for a privacy preserving feasibility query portal for distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) patient data repositories: design and implementation study. J Med Internet Res 2022. [doi: 10.2196/preprints.36709]
7. Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016 Oct;63:277-294 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.007] [Medline: 27507090]
8. Dobbins N, Spital C, Black R, Morrison J, de Veer B, Zampino E, et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. J Am Med Inform Assoc 2020 Jan 01;27(1):109-118 [FREE Full text] [doi: 10.1093/jamia/ocz165] [Medline: 31592524]
9. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009 Sep 01;16(5):624-630. [doi: 10.1197/jamia.m3191]
10. Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform 2018 Dec(2):1-10. [doi: 10.1200/cci.17.00067]
11. Forschungsnetz Covid-19 - SIMPLIFIER. URL: https://simplifier.net/ForschungsnetzCovid-19/~guides [accessed 2021-11-01]
12. Singh S, Aswal M. A state of the art on big data with semantic web technologies. In: Gaur L, Solanki A, Jain V, Khazanchi D, editors. Advances in Computer and Electrical Engineering. Hershey, PA: IGI Global; 2020:213.
13. CODEX - Gecco to UI and Mapper Internet. URL: https://github.com/num-codex/codex-gecco-to-ui-profiles [accessed 2021-11-03]
14. Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semantics 2018 Sep 17;9(1):24 [FREE Full text] [doi: 10.1186/s13326-018-0191-z] [Medline: 30223897]
15. Gulden C, Mate S, Prokosch H, Kraus S. Investigating the capabilities of FHIR Search for clinical trial phenotyping. In: German Medical Data Sciences: A Learning Healthcare System. Amsterdam: IOS Press; 2018:3-7.
16. Kiel A, Deppenwiese N, Kroll B, Engels C, Ebert L, Lablans M, et al. Feasibility studies with HL7 FHIR® and Clinical Quality Language. In: 16th Leipzig Research Festival for Life Sciences 2020. Germany: Faculty of Medicine, Leipzig University; 2020:105.
17. Hürsch W, Lopes C. Separation of Concerns. URL: https://www2.ccs.neu.edu/research/demeter/papers/publications-abstracts.html#SEP-CONCERNS [accessed 2022-04-06]
18. REDCap. URL: https://www.project-redcap.org/ [accessed 2021-11-25]
19. num-codex/odm2fhir. URL: https://github.com/num-codex/odm2fhir [accessed 2022-04-06]
20. num-codex/codex-testdata-to-sq. URL: https://github.com/num-codex/codex-testdata-to-sq [accessed 2022-04-06]
21. 7. SNOMED CT Expressions. URL: https://confluence.ihtsdotools.org/display/DOCSTART/7.+SNOMED+CT+Expressions [accessed 2022-04-06]
22. Observational Health Data Sciences and Informatics. ATLAS – A Unified Interface for the OHDSI Tools. URL: https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools/ [accessed 2022-04-06]
23. Maier C, Kapsner LA, Mate S, Prokosch H, Kraus S. Patient cohort identification on time series data using the OMOP common data model. Appl Clin Inform 2021 Jan 27;12(1):57-64 [FREE Full text] [doi: 10.1055/s-0040-1721481] [Medline: 33506478]
24. Karim MR, Nguyen BP, Zimmermann L, Kirsten T, Löbe M, Meineke F, et al. A Distributed Analytics Platform to Execute FHIR-Based Phenotyping Algorithms. URL: http://ceur-ws.org/Vol-2275/paper8.pdf [accessed 2022-04-06]
25. Gruendner J, Gulden C, Kampf M, Mate S, Prokosch H, Zierk J. A framework for criteria-based selection and processing of Fast Healthcare Interoperability Resources (FHIR) data for statistical analysis: design and implementation study. JMIR Med Inform 2021 Apr 01;9(4):e25645 [FREE Full text] [doi: 10.2196/25645] [Medline: 33792554]

26. Schüttler C, Prokosch H, Hummel M, Lablans M, Kroll B, Engels C, German Biobank Alliance IT development team. The journey to establishing an IT-infrastructure within the German Biobank Alliance. PLoS One 2021 Sep 22;16(9):e0257632 [FREE Full text] [doi: 10.1371/journal.pone.0257632] [Medline: 34551019]
27. Uciteli A, Beger C, Kirsten T, Meineke FA, Herre H. Ontological representation, classification and data-driven computing of phenotypes. J Biomed Semantics 2020 Dec 21;11(1):15 [FREE Full text] [doi: 10.1186/s13326-020-00230-0] [Medline: 33349245]
28. Blaze. URL: https://github.com/samply/blaze [accessed 2021-11-04]
29. Hauser R, Quine D, Ryder A, Campbell S. Unit conversions between LOINC codes. J Am Med Inform Assoc 2018 Feb 01;25(2):192-196 [FREE Full text] [doi: 10.1093/jamia/ocx056] [Medline: 28637208]
30. Benson T, Grieve G. Why interoperability is hard. In: Principles of Health Interoperability. Cham: Springer International; 2021:21-40.
31. Medizin Informatik Initiative. ABIDE_MI. URL: https://www.medizininformatik-initiative.de/de/use-cases-und-projekte/abidemi [accessed 2022-04-06]
32. Ganslandt T, Boeker M, Löbe M, Prasser F, Schepers J, Semler S, et al. Der Kerndatensatz der Medizininformatik-Initiativein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene. Forum der Medizin-Dokumentation und Medizin-Informatik 2018;20(1):21.
33. Lambarki M, Kern J, Croft D, Engels C, Deppenwiese N, Kerscher A, et al. Oncology on FHIR: a data model for distributed cancer research. In: Röhrig R, Beißbarth T, Brannath W, Prokosch HU, Schmidtmann I, Stolpe S, et al, editors. Studies in Health Technology and Informatics. Amsterdam: IOS Press; 2021.
34. openEHR. Archetype Query Language (AQL). URL: https://specifications.openehr.org/releases/QUERY/latest/AQL.html [accessed 2022-04-06]
35. Fette G, Kaspar M, Liman L, Ertl M, Krebs J, Dietrich G, et al. Query translation between AQL and CQL. Stud Health Technol Inform 2019 Aug 21;264:128-132. [doi: 10.3233/SHTI190197] [Medline: 31437899]

Abbreviations

CQL: Clinical Quality Language
DIC: Data Integration Centers
EHR: electronic health record
ETL: extract, transform, and load
FHIR: Fast Healthcare Interrogability Resources
GECCO: German Corona Consensus Dataset
ICD-10-GM: International Classification of Diseases and Related Health Problems, 10th edition, German version
JSON: JavaScript Object Notation
MII: Medical Informatics Initiative
PCE: postcoordinated expression
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms
SQ: Structured Query
SQL: Structured Query Language
UI: user interface

Edited by C Lovis; submitted 17.12.21; peer-reviewed by J Saß, M Kaspar; comments to author 07.01.22; revised version received 27.01.22; accepted 13.02.22; published 27.04.22.

Please cite as:
Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, Prokosch HU, Gruendner J
Generation of a Fast Healthcare Interoperability Resources (FHIR)-based Ontology for Federated Feasibility Queries in the Context of COVID-19: Feasibility Study
JMIR Med Inform 2022;10(4):e35789
URL: https://medinform.jmir.org/2022/4/e35789
doi:10.2196/35789
PMID:35380548

Original Paper

The Effect of an Additional Structured Methods Presentation on Decision-Makers' Reading Time and Opinions on the Helpfulness of the Methods in a Quantitative Report: Nonrandomized Trial

Jan Koetsenruijter[1], PhD; Pamela Wronski[1], MSc; Sucheta Ghosh[2], PhD; Wolfgang Müller[2], PhD; Michel Wensing[1], PhD

[1]Department of General Practice and Health Services Research, University Hospital Heidelberg, Heidelberg, Germany

[2]Scientific Databases and Visualization Group (SDBV), Heidelberg Institute for Theoretical Studies gGmbH, Heidelberg, Germany

Corresponding Author:
Jan Koetsenruijter, PhD
Department of General Practice and Health Services Research
University Hospital Heidelberg
Im Neuenheimer Feld 130.3
Heidelberg, 69120
Germany
Phone: 49 6221 56 4743
Email: jankoetsenruyter@hotmail.com

Abstract

Background: Although decision-makers in health care settings need to read and understand the validity of quantitative reports, they do not always carefully read information on research methods. Presenting the methods in a more structured way could improve the time spent reading the methods and increase the perceived relevance of this important report section.

Objective: To test the effect of a structured summary of the methods used in a quantitative data report on reading behavior with eye-tracking and measure the effect on the perceived importance of this section.

Methods: A nonrandomized pilot trial was performed in a computer laboratory setting with advanced medical students. All participants were asked to read a quantitative data report; an intervention arm was also shown a textbox summarizing key features of the methods used in the report. Three data-collection methods were used to document reading behavior and the views of participants: eye-tracking (during reading), a written questionnaire, and a face-to-face interview.

Results: We included 35 participants, 22 in the control arm and 13 in the intervention arm. The overall time spent reading the methods did not differ between the 2 arms. The intervention arm considered the information in the methods section to be less helpful for decision-making than did the control arm (scores for perceived helpfulness were 4.1 and 2.9, respectively, range 1-10). Participants who read the box more intensively tended to spend more time on the methods as a whole (Pearson correlation 0.81, *P*=.001).

Conclusions: Adding a structured summary of information on research methods attracted attention from most participants, but did not increase the time spent on reading the methods or lead to increased perceptions that the methods section was helpful for decision-making. Participants made use of the summary to quickly judge the methods, but this did not increase the perceived relevance of this section.

(JMIR Med Inform 2022;10(4):e29813) doi:10.2196/29813

KEYWORDS

decision-making; health care reports; reading behavior; research methods; eye-tracking; perceived importance; electronic health records; feasibility; quantitative methods

Introduction

Many quantitative reports are produced to help decision-makers in clinical practice, management, and health care policy. For the adequate use of these reports in decision-making, understanding the research methods used to generate their results and conclusions is essential for an assessment of the validity of the presented quantitative data. Previous studies have shown that the methodological limitations of claims made about health

interventions are neglected or not well understood by readers [1]. Even policy makers often fail to think critically about the trustworthiness of claims, and many people do not grasp that two things can be associated without one necessarily causing the other [1]. Various studies have shown that clinicians do not completely understand the information on treatment effects from meta-analyses [2] and that midwives and obstetricians are often unable to correctly interpret probabilistic screening information [3]. The QuantEV study examined the reading behavior of potential decision-makers by showing them a quantitative data report and found that reading behavior in the methods sections was variable and that overall, the section was not read thoroughly [4]. Critical assessment of the validity of data and methods is a part of the education of many health professionals and health care decision-makers (as one source puts it: "You cannot judge results without judging methods" [5]). Recently, an alliance of 24 researchers stated that teaching people to think critically about claims and comparisons will help them to make better decisions [1]. As more large-scale data is routinely collected and becomes available for decision-makers, the importance of knowing and understanding the validity and limitations of data reports has increased accordingly.

So far, a considerable body of research has focused on the evaluation and improvement of the critical appraisal skills of research users; in other words, how to train these users to increase their knowledge and improve their attitudes toward the use of research evidence [6]. Multiple studies have addressed the topic of reporting evidence (in this context, this corresponds to reporting results) and have given recommendations on how to present quantitative results visually [7,8]. By contrast, few intervention studies have focused on how research reports should present the validity of the evidence (ie, how they should present the methods). Moreover, many studies on reading behavior have used measurements that depend on self-reporting, or they have used methods like thinking aloud or the click-and-read method, all of which have uncertain validity [9-11]. Thus, we observe a need for readers to pay more attention to the methods used in reports and a lack of studies on how to achieve this. In the light of these observations, we developed an intervention which aimed at enhancing the understanding of the methods used in reports.

To design this intervention, we built on the results of a previous study (QuantEV), which assessed the reading behavior of future health policy decision-makers who were given quantitative data reports [4]. That study showed a high variation in reading time for the methods section, indicating that some decision-makers read the methods well, but others hardly paid attention to them. Reasons for paying less attention varied, but time constraints emerged as an important factor. In particular, the interplay between perceived relevance and time constraints led some participants to spend time on only those sections that they perceived as most relevant to decision-making. Although it is not easy to improve perceptions of relevance and methodological knowledge, time constraints can be addressed by reducing the time necessary to read the methods section.

To test whether reducing the time necessary to read the methods section would increase the attention paid to it, we developed an information textbox that offered a structured summary of a report's methods section. This textbox was placed at the beginning of the methods section in the upper left position, as this position attract readers' initial attention [12]. We designed the box while following guidelines developed for designing readable patient education materials. These guidelines state that critical information should be placed prominently, important elements and key points should be highlighted with visual cues (using devices such as boxes), and lists should be bulleted, so that they are easier to follow [13]. The textbox showed the structure of the report with headings that corresponded to the elements that readers usually find first during the skimming phase of reading [14]. We also used appropriate highlighting, which has been shown to enhance comprehension [15] and added tables, which are read more extensively than free text [16]. Drawing attention to a textbox enables readers to quickly judge whether there is anything questionable about the methods. In this way, adding a box can motivate readers to read more of the full methods section. Thus, this study set out to investigate whether adding a box led readers to read the methods more extensively. A secondary aim was to examine how this box influenced readers' appreciation of the importance of the methods for decision-making.

We hypothesized that by including a box with the highlights of the methods section, more participants would read at least a part of the methods section and the overall attention paid to the methods section would increase.

Methods

Study Design

We performed a nonrandomized pilot trial with a historical control arm in a computer laboratory setting. The trial used a computer-based quantitative data report; outcome measures were obtained with an eye tracker, a questionnaire, and a semistructured interview. The aim was to explore whether presenting the methods section in a structured and summarized manner could increase reading time and the perceived importance of the methods section.

Study Population and Research Setting

The study population was sampled from medical students in their seventh semester of study or later. All were potential future health care professionals who might become involved in local health care policy making. The eye-tracking measurements required that participants were not blind and did not have implanted artificial lenses. For the reading task, we requested that subjects have good knowledge of the German language. Students were invited to join the study with an email that was sent by the study program coordinators or the secretary. Additionally, posters were placed on campus and short presentations were given to students studying for bachelor's or master's degrees. For the control group, we used a historical control arm taken from the original project; these participants thus received the standard report. For the intervention group, a new sample was selected following the same criteria as the control group. A total sample size of at least 30 participants was felt to be sufficient for the exploratory nature of this pilot study.

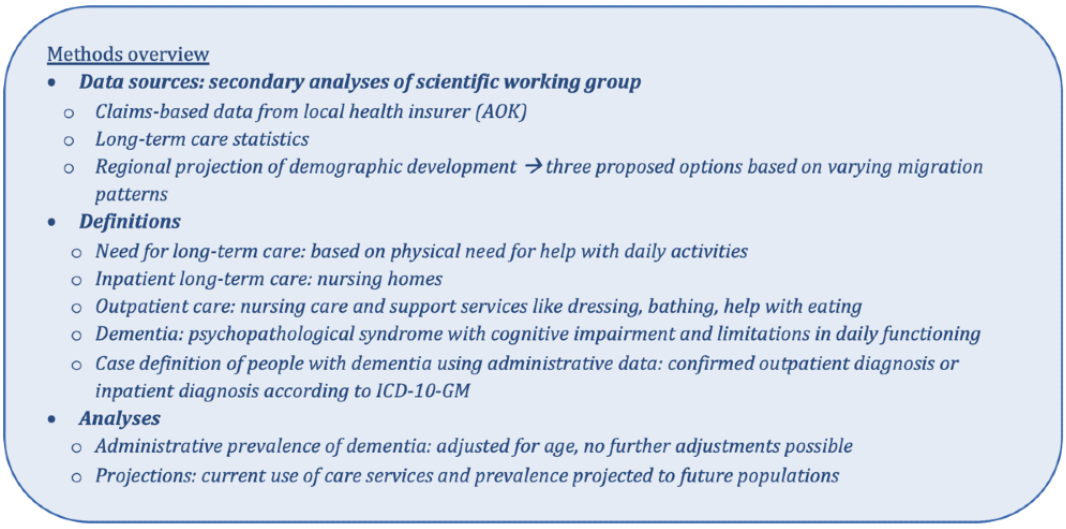
Data were collected at the eye-tracker laboratory of the Scientific Databases and Visualization group at the Heidelberg Institute for Theoretical Studies between April 2019 and March 2020.

Intervention

Participants were presented a decision scenario in the field of health care policy making with a quantitative data report to support the decision. The scenario was hypothetical, yet realistic: the participants had to advise the local district administrator on the use of additional funds for long-term care. The participants had to choose from 3 predefined options given by the research team. They were instructed to spend no more than 20 minutes on both reading the report and making the decision. We assumed a reading speed of 5 minutes per 1000 words, so given that the report contained around 4400 words, this time constraint was tight [17]. This was a deliberate choice, as in real decision-making scenarios, the amount of available information exceeds the time to read all of it. We also wanted to force the participants to restrict their reading time to the report sections that they found most valuable, thereby revealing how they prioritized the sections. The presented report was written in German and was 13 pages long, containing 3915 words. It was structured as a short project report, comprising a title page, a table of contents, an introduction (together approximately 1.5 pages), a methods section (approximately 3.5 pages), a results section (approximately 4.5 pages), and a discussion and conclusion section (together approximately 1 page). The quantitative data presented in the report were real descriptive figures on the current and projected demand and supply of long-term-care services in the region of interest; the data were based on secondary data analyses of real data [18].

The historical control arm received the original report with a traditional methods section, whereas the intervention arm received a version of the report with an additional textbox containing highlights of the methods. This box was developed to offer a structured presentation of the key aspects of the study design and methods. We designed the box to follow the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for conference abstracts, as provided by the Equator Network [19]. Thematically related bullet points were presented following the original structure of the full methods section (ie, data sources, definitions, and analyses) (Figure 1). Thus, this box provided the most relevant information at a glance and could help readers quickly grasp key information about the methods.

Figure 1. Example of textbox showing highlights of the study methods.



Measurements

We collected 3 different types of measurements from each participant: eye-tracking data during the performance of the task, answers to a questionnaire, and findings from a face-to-face, semistructured interview conducted after the task. For the eye tracking, we used the Tobii-X1 Light (Tobii AB) [20], a desktop-mounted, binocular eye tracker, and the Tobii eye-tracker software (version 3.4.8).

Based on the eye-tracking data, we extracted 3 measurements and calculated their mean values for the methods section (which included the box in the intervention group). First, we recorded the time spent (in minutes) to read the report and complete the task. Second, we computed the average fixation duration (in milliseconds) with the Tobii I-VT fixation filter and Tobii eye-tracker software [21]. Fixation duration was used as an indicator of attention to the report and information processing by the reader [22]. Fixation duration for individual readers varies during reading, and this variability can be used to provide a real-time reflection of ongoing cognitive and language processing [22]. Fixation duration serves as a primary source of evidence for testing theories of reading, and the central focus of computational models of skilled reading is to account for the influence of text processing on fixation duration [23-25]. Third, pupillary response was measured as an indicator of cognitive load [26-28]. The pupillary response was calculated as the change in pupil diameter from the average value as the eye passed over the screen [29]. While fixation duration accounts for cognitive and language processing, the pupillary response is associated with several cognitive functions, such as mental effort, interest, and decision-making [30-32]. All these cognitive functions, which cause variation in the size of the pupil, are related to attention [33,34]. Our main hypothesis was that the time spent on reading the methods section would be higher in

the intervention group. Furthermore, we expected that attention would be higher in the intervention group.

The questionnaire was used to collect individual characteristics, such as age, sex, and the participants' risk literacy. Risk literacy was measured using the validated Berlin numeracy test, which is a 4-item paper and pencil test taken in the German language [35]. We expected that participants with higher risk literacy would have more affinity with the quantitative methods. Risk literacy was therefore measured to control for potential differences between study arms. Finally, the participants were asked to assess each section of the report (ie, the introduction, methods, results, discussion, and conclusion) for understandability and helpfulness during the decision-making task, both on a 10-point Likert scale.

A semistructured question guide was developed to explore the experiences of the participants in completing the task. Interview questions were developed by the authors in cooperation with 2 colleagues with a sociology and health science background. The participants were encouraged to speak frankly about their experiences generally and about the way they had read the report specifically. The interviews were audio recorded and transcribed. Additional details on these methods have been previously published [4].

Analyses

Participants were the unit of analysis. Questionnaire and eye-tracking data (after data preparation with the Tobii eye-tracker software) were analyzed using SPSS Statistics (version 25; IBM). Descriptive analyses and the *t* test were used to find differences between the study arms. Additionally, the Pearson correlation was calculated to explore the relationships between fixation, pupillometric data, and participant characteristics. Considering the exploratory nature of the study and the small sample size, we regarded a P value <.1 as significant. For analyzing the interviews, qualitative content analysis was conducted to explore reasons mentioned by participants as to why they gave more or less attention to a report section during decision-making. The qualitative content analysis was conducted by 2 of the authors with the support of a third colleague.

Ethics Approval

This study was approved by the research ethics committee of Heidelberg University Hospital (S-857/2018) and was part of the QuantEV project, which has been described previously [4]. Before data collection, participants were informed orally and in writing about the study context, the data collection procedure, and data security. All participants provided consent for participation. Participation was voluntary, and study withdrawal was possible at any time before the collected data were anonymized.

Results

In total, 35 participants were included, 22 in the control arm and 13 in the intervention arm (Table 1). In both the control and intervention arms, women were more represented (18/22, 82%; and 8/13, 64%, respectively). The average age of the participants was 23.7 years; this was similar between the arms. The average risk literacy score was 68.0 (ie, 68% of the answers were correct); this was similar between the control and intervention arms (with an average risk literacy score of 69.3 and 65.4, respectively). The decisions made after reading the report were also similar in both groups, with option C (increasing ambulant nursing capacity) being the dominant choice, chosen by 27 of 35 (77%) of the participants.

Table 1. Characteristics of study participants.

Characteristics	Control group (n=22)	Intervention group (n=13)	Overall (N=35)
Female, n (%)	18 (82)	8 (64)	27 (76)
Age (years), mean (SD)	23.9 (1.5)	23.5 (2.3)	23.7 (1.8)
Risk literacy score, mean (SD)	69.3 (33.6)	65.4 (31.5)	68.0 (32.4)
Decision on how to spend funds			
Option A: more support for informal caregivers, n (%)	2 (9)	2 (15)	4 (11)
Option B: more nursing home capacity, n (%)	3 (14)	1 (8)	4 (11)
Option C: more ambulant nursing capacity, n (%)	17 (77)	10 (77)	27 (77)

Table 2 shows the eye-tracking and questionnaire results for participants from both study arms and eye-tracking data for the intervention group only. Differences between the study arms were tested for statistical significance (rightmost column). The overall time spent on the methods and the pupillary response did not differ between the 2 study arms. The average fixation duration was higher in the control arm, but not significantly (0.445 seconds vs 0.337 seconds, P=.56). The questionnaire results did not show a difference in perceived understandability, but the intervention group found the methods less helpful for decision-making (score 2.9 vs 4.1, P=.09). These findings do not provide support for the hypothesis that the box would increase attention paid to the methods section. The box-only results for the intervention group showed that the participants spent about half a minute reading the box, while the other eye-tracking values were similar to those for the overall methods.

Figure 2 shows box plots for time spent on reading the methods section in both study arms. Whereas the mean time was very similar, the median time, as well as the 25th and 75th percentiles, were lower in the intervention arm. The variation was also higher in the intervention group, mostly due to 2 participants who spent over 10 minutes reading the methods.

Our findings for variation in time spent looking at the box showed that 2 participants hardly looked at it at all, while the other participants spent between 0.2 and 1 minute looking at it. Most of the participants took at least a brief look at the box, which could provide support for our hypothesis that the box would lead to more participants reading at least a minimal part of the methods section.

Figure 3 shows the relationship between the time spent reading the box and the time spent subsequently reading the whole methods section in the intervention group. Each data point represents a single participant. There was a clear positive relationship between the 2 parameters: participants who spent more time reading the box tended to spend more time reading the full methods (Pearson correlation 0.81, *P*=.001). This relationship held for both the single participant who did not read either the box or the full methods and for the 3 participants who spent the most time on the box, who were also the ones who spent the most time on the methods.

We used heat maps to perform a qualitative analysis of reading patterns. This confirmed the relationship between time spent reading the box and the full methods (Figure 4). Participants who read the box spent more time reading the methods section. In some cases, a participant only briefly skimmed the box and ignored the rest of the methods. This provides support for the idea that introducing the box prompted participants to read at least a minimal amount of the methods section. There was no case in which a participant read the box thoroughly and then skipped reading the full methods. This supports the idea that the box could potentially increase overall attention paid to the methods section. Some participants read the whole methods section, including the box, while others paid more attention to specific subsections.

Table 2. Eye-tracking and questionnaire measures by study arm.

Measure	Control	Intervention		*P* value Δ: control – intervention
	Overall	Overall	Box only	
Eye tracking				
Time spent in minutes, mean (SD)	4.03 (2.39)	4.07 (3.68)	0.46 (0.31)	.96
Average fixation duration in seconds, mean (SD)	0.445 (0.643)	0.337 (0.213)	0.316 (0.145)	.56
Pupillary response in mm, mean (SD)	0.033 (0.011)	0.034 (0.017)	0.032 (0.014)	.91
Questionnaire				
Understandability, range 1-10, mean (SD)	6.6 (2.1)	6.6 (2.0)	N/A	.96
Helpfulness for decision-making, range 1-10, mean (SD)	4.1 (2.2)	2.9 (1.3)	N/A	.09

[a]N/A: not applicable.

Figure 2. Time spent reading the methods section in reports with and without an added box (in the intervention and control groups) and time spent reading the box itself (in the intervention group) in minutes.

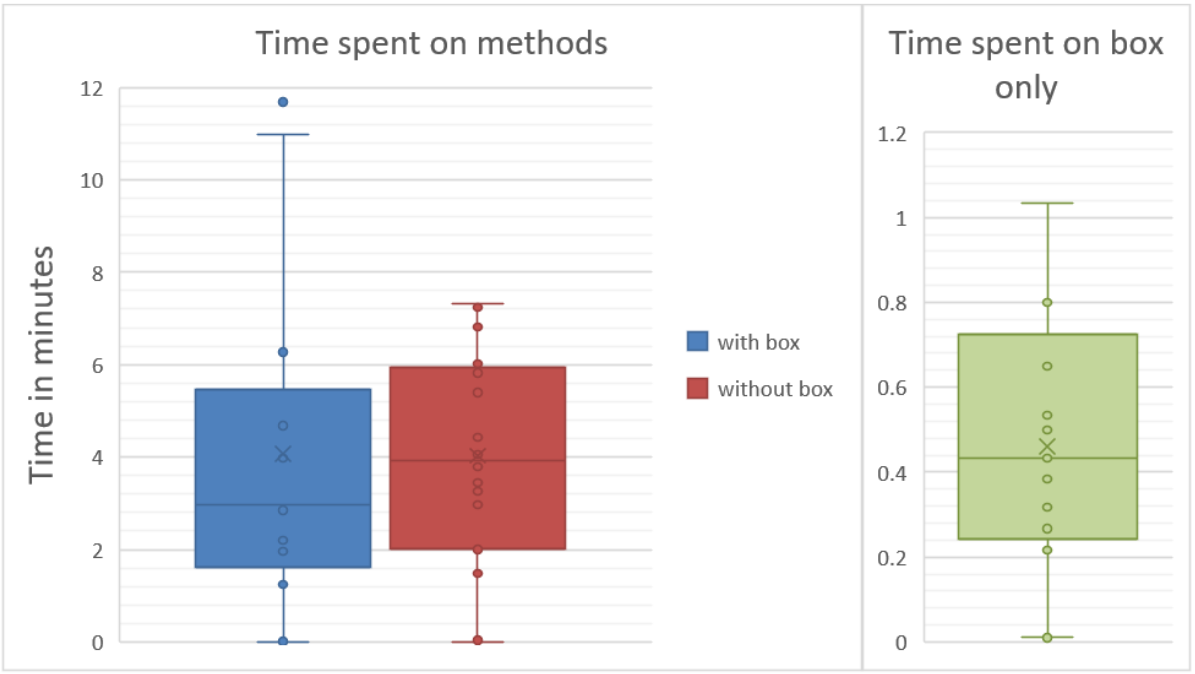


Figure 3. Time spent on reading the box vs the full methods.

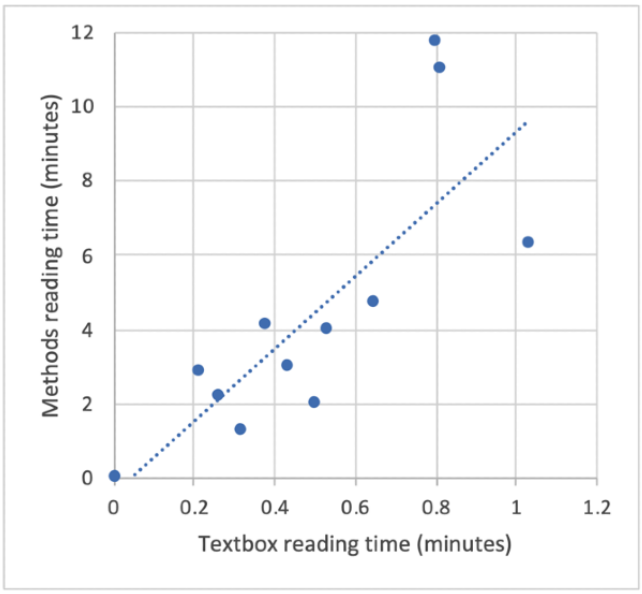
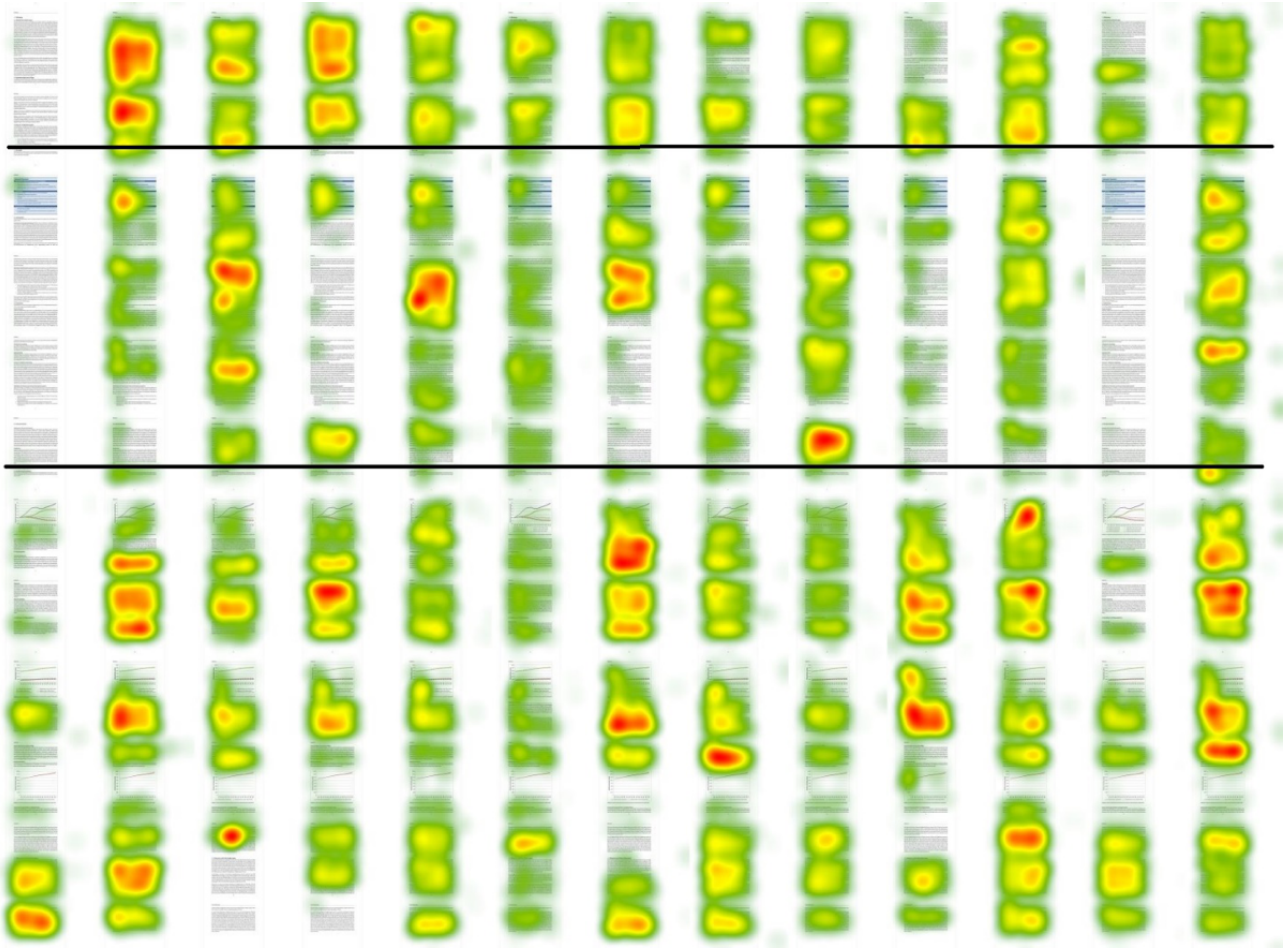


Figure 4. Heat map of intervention group reading time. The methods section is shown between the horizontal lines.



In the qualitative interviews, participants reported why they preferred specific report sections. Some participants reported that they paid less attention to the methods because they did not perceive them as relevant for decision-making, and that they trusted the authors to have used valid methods. Other participants reported that they used the box to gain a broad understanding of the methods, and then only briefly scanned the full section, because they did not perceive it as very important and felt it was not necessary to read it thoroughly.

Discussion

Adding a textbox with a structured summary of the methods did not increase the total time spent reading the full methods

section, but it was successful in attracting attention, as most participants at least skimmed the box. However, including the box resulted in a lower appreciation of the helpfulness of the information on research methods. Participants who spent more time on the box also spent more time on the methods in general. The finding that the box seemed to attract attention provides support for our hypothesis that it led more participants to read at least a minimal portion of the methods section. Our findings from the heat map also support this hypothesis. However, as overall reading time did not increase, and we even found that appreciation of the helpfulness of the methods section decreased, we did not obtain support for our hypothesis that the box would increase the overall attention paid to the methods section.

Hypothetically, including the box could have either enhanced or reduced time spent reading the methods section. If it had been seen as complementary information, as we hypothesized it would be, it could have motivated participants to read the full text. However, the finding that there was no increase in overall reading time for the methods does not support the idea that the box was complementary. The linear relationship between time spent reading the box and the methods could have been caused by general interest, or lack thereof, in the methods, rather than indicating that participants read the box purposefully, to quickly gain an overview of the methods without having to read the full section. A potential explanation was provided by the interviews: some participants indicated that they paid less attention to the methods because, considering the time constraints, they did not perceive this section as relevant. The specific sample of medical students examined in this study could also have been a factor, as they might not have been very comfortable with quantitative and statistical methods [36]. Thus, rather than serving a complementary function, the box could have been used as a substitution for the full text, resulting in less attention paid to the methods as a whole. This explanation is supported by the findings that participants did not spend less time on the methods overall and that there was a positive, linear relationship between time spent reading the box and the methods.

The finding that fixation duration was shorter in the intervention group (although this was nonsignificant) could indicate that there was less engagement with the presented text [22]. This explanation would correspond with the finding that by adding the box, the methods were perceived as less useful to complete the presented task. However, fixation duration could also indicate language processing [37]. If the box helped the reader to become familiar with the topic, the full methods section might have been perceived as less complex, reducing the need for language processing and enabling an increase in reading speed. Our use of eye tracking in addition to a questionnaire allowed us to collect rich data on reading behavior that was not influenced by the limitations of self-reported behavior. Limitations in our study were caused by the specific participants we recruited and our measurements of reading time. Our sample consisted of future health care professionals with only limited experience in decision-making, meaning that the findings may not be fully generalizable to more experienced policy makers, who might have perceived and used the report differently. However, our participants were already advanced students and all had received training in interpreting studies, reflected by a slightly higher numeracy level than general practitioners and other medical students [38]. Our method of measuring time spent on the methods did not automatically mean that a subject also read the text. Nevertheless, our findings on average fixation duration suggest that our subjects did read the text, as fixation is, on average, about 0.25 seconds while reading [22]. Additionally, the study design was primarily tailored to the design of a past project rather than to the present intervention study.

In this study, we aimed to explore whether presenting the methods of a report as a structured summary could increase time spent reading the methods section. Our findings indicate that including a box might help to attract attention, but that it might not increase overall interest in the methods section. The intervention might have motivated more decision-makers to read at least some of the methods and helped them judge if the methods needed a full inspection. However, the limited attention paid to the methods by some participants, who considered the methods not relevant for decision-making, is a problem that might not be solvable by changing the input (ie, the format of the report). Rather, it might require an intervention at the individual level to increase awareness of the relevance of the methods section to decision-making.

Acknowledgments

The authors acknowledge and thank the study participants for their time and contributions. This study was funded by Klaus Tschira Stiftung gGmbH (project number 00.349.2018). The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Data Availability

The data sets used and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

JK and PW conceptualized the manuscript, with JK taking the lead on writing of all drafts, integrating feedback upon reviews, and finalizing the manuscript. SG and WM led development of the eye-tracking experiment. JK, PW, and SG analyzed the data. MW and WM developed the project in which the study was embedded and secured funding. All authors contributed to the final manuscript, and they read and approved it prior to submission.

Conflicts of Interest

None declared.

References

1. Aronson JK, Barends E, Boruch R, Brennan M, Chalmers I, Chislett J, et al. Key concepts for making informed choices. Nature 2019 Aug;572(7769):303-306. [doi: 10.1038/d41586-019-02407-9] [Medline: 31406318]
2. Johnston BC, Alonso-Coello P, Friedrich JO, Mustafa RA, Tikkinen KA, Neumann I, et al. Do clinicians understand the size of treatment effects? A randomized survey across 8 countries. CMAJ 2016 Jan 05;188(1):25-32 [FREE Full text] [doi: 10.1503/cmaj.150430] [Medline: 26504102]
3. Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. BMJ 2006 Aug 05;333(7562):284 [FREE Full text] [doi: 10.1136/bmj.38884.663102.AE] [Medline: 16840441]
4. Wronski P, Wensing M, Ghosh S, Gärttner L, Müller W, Koetsenruijter J. Use of a quantitative data report in a hypothetical decision scenario for health policymaking: a computer-assisted laboratory study. BMC Med Inform Decis Mak 2021 Jan 28;21(1):32 [FREE Full text] [doi: 10.1186/s12911-021-01401-4] [Medline: 33509172]
5. Yudkin B. Critical reading: making sense of research papers in life sciences and medicine. London: Routledge; Apr 2006.
6. Young T, Rohwer A, Volmink J, Clarke M. What are the effects of teaching evidence-based health care (EBHC)? Overview of systematic reviews. PLoS One 2014;9(1):e86706 [FREE Full text] [doi: 10.1371/journal.pone.0086706] [Medline: 24489771]
7. Gerrits RG, Kringos DS, van den Berg MJ, Klazinga NS. Improving interpretation of publically reported statistics on health and healthcare: the Figure Interpretation Assessment Tool (FIAT-Health). Health Res Policy Syst 2018 Mar 07;16(1):20 [FREE Full text] [doi: 10.1186/s12961-018-0279-z] [Medline: 29514711]
8. Petkovic J, Welch V, Jacob MH, Yoganathan M, Ayala AP, Cunningham H, et al. The effectiveness of evidence summaries on health policymakers and health system managers use of evidence from systematic reviews: a systematic review. Implement Sci 2016 Dec 09;11(1):162 [FREE Full text] [doi: 10.1186/s13012-016-0530-3] [Medline: 27938409]
9. Ummelen N, Neutelings R. Measuring reading behavior in policy documents: a comparison of two instruments. IEEE Trans. Profess. Commun 2000;43(3):292-301. [doi: 10.1109/47.867945]
10. Tenopir C, King DW, Clarke MT, Na K, Zhou X. Journal reading patterns and preferences of pediatricians. J Med Libr Assoc 2007 Jan;95(1):56-63 [FREE Full text] [Medline: 17252067]
11. Guan Z, Lee S, Cuddihy E, Ramey J. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. : Association for Computing Machinery; 2006 Apr Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06); April 22-27, 2006; Montréal Québec Canada p. 1253-1262. [doi: doi.org/10.1145/1124772.1124961]
12. Holmqvist K, Wartenberg C. The Role of Local Design Factors for Newspaper Reading Behaviour: An Eye-Tracking Perspective. Lund University Cognitive Studies 2005;127:1-21.
13. Aldridge MD. Writing and designing readable patient education materials. Nephrol Nurs J 2004;31(4):373-377. [Medline: 15453229]
14. Day T. Success in Academic Writing. London: Red Globe Press; 2018.
15. Beymer D, Russell D, Orton P. An Eye Tracking Study of How Font Size and Type Influence Online Reading. 2008 Sep Presented at: People and Computers XXII Culture, Creativity, Interaction (HCI); 1 - 5 September 2008; Liverpool, UK p. 15-18. [doi: 10.14236/ewic/HCI2008.23]
16. de Kock E, van Biljon J, Pretorius M. Usability evaluation methods: Mind the gaps. : Association for Computing Machinery Presented at: SAICSIT '09: Proceedings of the 2009 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists; October 12 - 14, 2009; Vanderbijlpark Emfuleni, South Africa p. 122-131. [doi: 10.1145/1632149.1632166]
17. Rubin GS, Turano K. Reading without saccadic eye movements. Vision Research 1992 May;32(5):895-902. [doi: 10.1016/0042-6989(92)90032-e]
18. Modellprojekt Sektorenübergreifende Versorgung in Baden-Württemberg - Projektbericht. 2018. URL: https://sozialministerium.baden-wuerttemberg.de/de/gesundheit-pflege/medizinische-versorgung/sektorenuebergreifende-versorgung/ [accessed 2021-09-12]
19. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007 Oct 20;370(9596):1453-1457. [doi: 10.1016/S0140-6736(07)61602-X] [Medline: 18064739]
20. Tobii TAB. No Tobii X2-30 Eye Tracker User's manual. 2014.
21. Olsen A, Matos R. Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. Association for Computing Machinery: ACM Press; 2012 Mar Presented at: ETRA '12: Eye Tracking Research and Applications; March 28 - 30, 2012; Santa Barbara, California p. 317-320. [doi: 10.1145/2168556.2168625]
22. Rayner K. Eye movements and attention in reading, scene perception, and visual search. Q J Exp Psychol (Hove) 2009 Aug;62(8):1457-1506. [doi: 10.1080/17470210902816461] [Medline: 19449261]

23. Breen M, Clifton C. Stress matters: Effects of anticipated lexical stress on silent reading. Journal of Memory and Language 2011 Feb;64(2):153-170. [doi: 10.1016/j.jml.2010.11.001]
24. Nuthmann A, Henderson JM. Using CRISP to model global characteristics of fixation durations in scene viewing and reading with a common mechanism. Visual Cognition 2012 Apr;20(4-5):457-494. [doi: 10.1080/13506285.2012.670142]
25. Engbert R, Nuthmann A, Richter EM, Kliegl R. SWIFT: A Dynamical Model of Saccade Generation During Reading. Psychological Review 2005;112(4):777-813. [doi: 10.1037/0033-295x.112.4.777]
26. Hartmann M, Fischer M. Pupillometry: The Eyes Shed Fresh Light on the Mind. Current Biology 2014 Mar;24(7):R281-R282. [doi: 10.1016/j.cub.2014.02.028]
27. Sweller J, Van Merrienboer JG, Paas FGWC. Cognitive Architecture and Instructional Design. Educational Psychology Review 1998;10:251-296. [doi: 10.1023/A:1022193728205]
28. Szulewski A, Roth N, Howes D. The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians. Academic Medicine 2015;90(7):981-987. [doi: 10.1097/acm.0000000000000677]
29. Attard-Johnson J, Ó Ciardha C, Bindemann M. Comparing methods for the analysis of pupillary response. Behav Res 2018 Oct 15;51(1):83-95. [doi: 10.3758/s13428-018-1108-6]
30. Van Slooten JC, Jahfari S, Knapen T, Theeuwes J. How pupil responses track value-based decision-making during and after reinforcement learning. PLoS Comput Biol 2018 Nov 30;14(11):e1006632. [doi: 10.1371/journal.pcbi.1006632]
31. Kahneman D, Beatty J. Pupil Diameter and Load on Memory. Science 1966 Dec 23;154(3756):1583-1585. [doi: 10.1126/science.154.3756.1583]
32. Krugman HE. Some Applications of Pupil Measurement. Journal of Marketing Research 1964 Nov;1(4):15. [doi: 10.2307/3150372]
33. Hoeks B, Levelt WJM. Pupillary dilation as a measure of attention: a quantitative system analysis. Behavior Research Methods, Instruments, & Computers 1993 Mar;25(1):16-26. [doi: 10.3758/bf03204445]
34. Wierda SM, van Rijn H, Taatgen NA, Martens S. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proc Natl Acad Sci USA 2012 May 29;109(22):8456-8460 [FREE Full text] [doi: 10.1073/pnas.1201858109] [Medline: 22586101]
35. Cokely E, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring risk literacy: The berlin numeracy test. Judgment and Decision Making 2012;7(1):25-47. [doi: 10.1037/t45862-000]
36. Herman A, Notzer N, Libman Z, Braunstein R, Steinberg DM. Statistical education for medical students--concepts are what remain when the details are forgotten. Stat Med 2007 Oct 15;26(23):4344-4351. [doi: 10.1002/sim.2906] [Medline: 17487940]
37. Henderson JM, Choi W, Luke SG, Desai RH. Neural correlates of fixation duration in natural reading: Evidence from fixation-related fMRI. NeuroImage 2015 Oct;119:390-397. [doi: 10.1016/j.neuroimage.2015.06.072]
38. Friederichs H, Birkenstein R, Becker JC, Marschall B, Weissenstein A. Risk literacy assessment of general practitioners and medical students using the Berlin Numeracy Test. BMC Fam Pract 2020 Jul 14;21(1):143 [FREE Full text] [doi: 10.1186/s12875-020-01214-w] [Medline: 32664885]

Edited by C Lovis; submitted 21.04.21; peer-reviewed by J Moll, I Wilson; comments to author 25.09.21; revised version received 20.12.21; accepted 31.01.22; published 12.04.22.

Please cite as:
Koetsenruijter J, Wronski P, Ghosh S, Müller W, Wensing M
The Effect of an Additional Structured Methods Presentation on Decision-Makers' Reading Time and Opinions on the Helpfulness of the Methods in a Quantitative Report: Nonrandomized Trial
JMIR Med Inform 2022;10(4):e29813
URL: https://medinform.jmir.org/2022/4/e29813
doi:10.2196/29813
PMID:35412464

Original Paper

Exploring Patient Multimorbidity and Complexity Using Health Insurance Claims Data: A Cluster Analysis Approach

Anna Nicolet[1*], PhD; Dan Assouline[1*], PhD; Marie-Annick Le Pogam[1], MD, MPH; Clémence Perraudin[1], PhD; Christophe Bagnoud[2], PhD; Joël Wagner[3], PhD, Prof Dr; Joachim Marti[1*], PhD, Prof Dr; Isabelle Peytremann-Bridevaux[1*], MD, MPH, DSc, Prof Dr

[1]Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland

[2]Groupe Mutuel, Martigny, Switzerland

[3]Department of Actuarial Science, Faculty of Business and Economics, and Swiss Finance Institute, University of Lausanne, Lausanne, Switzerland

[*]these authors contributed equally

Corresponding Author:
Anna Nicolet, PhD
Center for Primary Care and Public Health (Unisanté)
University of Lausanne
Route de la Corniche
Lausanne, 1010
Switzerland
Phone: 41 21 314 23 4
Email: anna.nicolet@unisante.ch

Abstract

Background: Although the trend of progressing morbidity is widely recognized, there are numerous challenges when studying multimorbidity and patient complexity. For multimorbid or complex patients, prone to fragmented care and high health care use, novel estimation approaches need to be developed.

Objective: This study aims to investigate the patient multimorbidity and complexity of Swiss residents aged ≥50 years using clustering methodology in claims data.

Methods: We adopted a clustering methodology based on random forests and used 34 pharmacy-based cost groups as the only input feature for the procedure. To detect clusters, we applied hierarchical density-based spatial clustering of applications with noise. The reasonable hyperparameters were chosen based on various metrics embedded in the algorithms (out-of-bag misclassification error, normalized stress, and cluster persistence) and the clinical relevance of the obtained clusters.

Results: Based on cluster analysis output for 18,732 individuals, we identified an outlier group and 7 clusters: individuals without diseases, patients with only hypertension-related diseases, patients with only mental diseases, complex high-cost high-need patients, slightly complex patients with inexpensive low-severity pharmacy-based cost groups, patients with 1 costly disease, and older high-risk patients.

Conclusions: Our study demonstrated that cluster analysis based on pharmacy-based cost group information from claims-based data is feasible and highlights clinically relevant clusters. Such an approach allows expanding the understanding of multimorbidity beyond simple disease counts and can identify the population profiles with increased health care use and costs. This study may foster the development of integrated and coordinated care, which is high on the agenda in policy making, care planning, and delivery.

(JMIR Med Inform 2022;10(4):e34274) doi:10.2196/34274

KEYWORDS

multimorbidity; pharmacy cost groups; cluster analysis; claims data; patient complexity; health claims; informatics

Introduction

Health care systems worldwide are facing considerable challenges from the increasing number of chronic and multimorbid patients, characterized by complex needs and frequent transitions between care settings [1]. In Switzerland, 2.2 million people report a chronic disease and nearly 20% of the population older than 50 years have multiple chronic diseases (multimorbidity) [2]. Although the trend of progressing multimorbidity is widely recognized [3-6], it is still unclear how best to take care of patients with multimorbidity and which interventions would be effective. For more than two decades, integrated and coordinated care have been developed worldwide [7]. Nevertheless, integrated and coordinated care faces continuing challenges such as scaling-up, implementation, and sustainability difficulties. Additionally, integrated and coordinated care requires development of novel approaches to evaluate and measure patients multimorbidity and complexity. This is key to stratify the targeted population and adapt the intervention to the needs of the patients. Often, such evaluations and measures rely on morbidity indices (eg, Charlson and Elixhauser) or on the number of (self-reported) chronic conditions or comorbidities [8]. Whereas the former were developed in an inpatient setting as predictors of mortality, the latter may not comprehensively reflect the patient's disease burden and complexity. Despite these limitations, they remain often used because of their relative accessibility and simplicity. In settings where electronic medical (health) records, national disease registries, or data on chronic conditions are unavailable, administrative health insurance claims data represent a potentially useful source of information. In fact, they are increasingly used in health services research, especially to express multimorbidity using pharmacy-based cost groups (PCGs) [9,10]. PCGs, based on use of prescribed drugs rather than on clinical information, were developed as a proxy for morbidity measure [11]. Although the approach has limitations related to underestimation of medicines used, unclaimed, or paid out-of-pocket and thus not present in the data or the assumption that the drug is used exclusively for treating the particular condition [11,12], it allows mapping patient profiles to reflect their morbidity status. As such mapping approaches and comorbidity counts are considered simplistic [13], researchers may consider alternative methods to investigate patient complexity more exhaustively. One such method is cluster analysis, which relies on the idea that many common conditions cluster together in the population in predictable patterns [13]. It has been shown that cluster analysis of real-world data for drug use research can be used for detecting clinically plausible subgroups [14]. Similar approaches of classifications based on multimorbidity patterns have been applied in the literature [14-16], but using PCGs as the multimorbidity indicator for cluster analysis is novel. In that context, the aim of our study is to investigate patient multimorbidity and complexity beyond simple mapping and counts of PCGs, using clustering methodology in claims data of Swiss residents aged ≥50 years.

Methods

Data Source and Sample

We included data of 240,511 insured people aged ≥50 years continuously enrolled in one of the largest health insurance companies in Switzerland, Groupe Mutuel, for the 2015-2018 period. In addition to demographic information (age and gender), data contained PCGs for each individual, costs covered by the patient (cost sharing), type of health insurance model (with or without gatekeeping), and reimbursed health care services: number of visits to various physicians with associated costs and physicians' specialization and hospitalizations. To identify insured persons with cost-intensive, chronic diseases and correspondingly high health care use based on their drug consumption, health insurance companies are translating the drug use data reflecting active ingredient and quantity, based on Anatomical Therapeutic Chemical and defined daily dose, into the PCGs. This procedure was developed and officially accepted by the Federal Office of Public Health in Switzerland [17]. In our study, the patients were classified as multimorbid when they were assigned two or more PCGs, based on their yearly drug use.

Ethical Considerations

Data were deidentified by the insurance company to guarantee anonymization, and ethical approval for this study was waived by the Cantonal Commission for the Ethics of Research on Human Beings (Lausanne, Switzerland).

Cluster Analysis

We adopted a clustering methodology based on random forests (RFs) [18]—a popular classification and regression tree-based method—that includes several steps and machine learning algorithms [19-21]. The methodology is inspired by a clustering methodology designed by Breiman and Cutler [19], the creators of RFs [20,21].

In a preprocessing step, we extracted 34 PCGs as the only input feature for the clustering procedure. We grouped the 34 PCGs into 15 disease categories, which were valued meaningful from a clinical perspective (Multimedia Appendix 1). We then considered the first year of information only, and extracted a 10% random sample, to allow for effective processing for the computationally expensive steps. To confirm the results, the random sampling was performed multiple times, which led to similar clusters. Finally, we discarded points showing no PCG or only one type of PCG. Since we ultimately use an algorithm to detect clusters based on density given by the distances between points, the presence of many identical points at the same positions may perturb the algorithm and unnecessarily make the computation more expensive. Keeping a small random sample of these points would reduce the perturbation but not change the results while adding a dispensable complication, notably for the hyperparameter selection needed to detect these additional clusters.

To initiate the clustering procedure, we created a synthetic data set of the same size as the original data, by random sampling from the distributions of each input variable within the data. The idea is then to train an RF model to classify synthetic and

original points, with the aim of taking advantage of the *proximity* measure, an embedded RF metric of similarity between points. An RF aggregates the prediction of multiple decision trees (DTs) by considering the class they predict in majority. DTs are classification models that separate the data points into subspaces (leaves) by imposing thresholds on the input variables and predicting the class within each subspace as the majority class. The proximity between two points is then computed as the number of times they fall in the same leaf across the trees in the forest. To stabilize the random effects of RFs, we trained 10 RF models, computed the proximities for all pairs of points for each model, and averaged them to obtain a mean proximity matrix characterizing the data. We then used multidimensional scaling (MDS) [22] to project the corresponding distance matrix (1 – proximity matrix / (number of trees)) in 2D while preserving the distances and allow for visualization of the resulting clusters. Finally, we applied hierarchical density-based spatial clustering of applications with noise (HDBSCAN) [23] to detect clusters within the obtained 2D data, after discarding the synthetic points from the data. HDBSCAN extracts clusters as dense gatherings of points separated by sparse regions with few points. Given that no cross-validation is possible with clustering methodologies, reasonable hyperparameters were chosen for the RF, MDS, and HDBSCAN steps based on various metrics embedded in the algorithms and the clinical relevance of the obtained clusters. The metrics includes the *out-of-bag* (*OOB*) misclassification error, which shows how well RF differentiates the original data from the synthetic one. The outcome reflects how much structure there is in the data [19]. Another metric was *normalized stress*, measuring whether the distances between points are reasonably preserved after projection [22], and the *cluster persistence*, HDBSCAN embedded metrics indicating how well the clusters are defined and separated from each other [23]. In practice, we used the HDBSCAN and Scikit-learn libraries (in Python) for the final clustering and all previous steps.

Results

After discarding individuals with missing information, our data set comprised 18,732 individuals (*points*). An initial examination of the data set exhibited three large "single" clusters that we extracted prior to the clustering procedure, showing no PCGs, only hypertension PCGs, and only mental disease PCGs, representing 67.9% (n=12,720), 9.7% (n=1813), and 4.1% (n=765) of the population, respectively. Clustering analyses, performed on the remaining 3434 patients not included in the latter "single" clusters, identified four distinct clusters: Cluster 0 to Cluster 3, numbered in the order in which they are detected while applying HDBSCAN (Figure 1). The clusters can be clearly visualized from this tree (Figure 2); and a good persistence of 0.29, 0.24, 0.15, and 0.24, respectively, was found. The average OOB misclassification error from the 10 RFs was 0.51, which is quite high, showing that RF does not differentiate well between the original and the synthetic data, and there is not much structure in the data. Regarding the performed MDS, the normalized stress was 0.31, indicating reasonable preserving of the distances between points.

The 4 detected clusters encompass different mixes of PCGs (Table 1 and Figure 3): Cluster 0 comprises a large mix of PCGs (mental + hypertension + pain + asthma [chronic obstructive pulmonary disease]) often appearing jointly; Cluster 1 comprises PCGs (thyroid, hypertension, glaucoma, and mix of others) appearing jointly less often; Cluster 2 comprises asthma, Parkinson, cardiac diseases, and pain rarely appearing jointly; and Cluster 3 comprises a large mix of PCGs almost never appearing jointly (single diseases).

The following description and interpretation of clusters is based on the descriptive statistics of health care use and costs data (Table 1), which help to understand the underlying principle of grouping individuals into PCG clusters. First, the members of Cluster 0 (n=817, 4.4%) had the highest number of PCGs and highest costs and health care use, and were referred to as "complex high-cost high-need patients" (for a detailed description, see Table 1). The degree of complexity in these settings was reflected as the combination of the following characteristics interpreted from descriptive statistics (Table 1): average number of PCGs, percentage of multimorbid patients, levels of health care use (eg, number of doctor consultations and hospital stays), and costs in the population subgroup. The members of Cluster 1 (n=709, 3.8%), although having multiple PCGs, had health care costs and use lower than in Cluster 0; thus, they were referred to as "slightly complex with inexpensive low-severity PCGs." The members of Cluster 2 (n=531, 2.8%) were of the oldest age and presented especially high use of hospitalizations and visits to the generalist doctor and, thus, were referred to as "oldest at high risk." High risk, interpreted in these settings from the descriptive statistics, was reflected by relatively high use of hospital care, yet lower than in the most complex cluster: long length of stay (5.6 and 6.6 nights for clusters "Oldest at risk" and "Complex high-cost high-need," respectively) and high inpatient costs (CHF 2749 [US $2950] and CHF 3109 [US $3333], respectively). The members of Cluster 3 (n=1056, 5.6%) were characterized by a relatively small number of PCGs (close to 1) and the highest costs of medications and, thus, were referred to as "patients with 1 costly disease."

Figure 1. MDS projection of the data in two dimensions. The four clusters found by HDBSCAN are marked by the different colors and coded with the labels 0, 1, 2, and 3. The code –1 refers to the outliers. HDBSCAN: hierarchical density-based spatial clustering of applications with noise; MDS: multidimensional scaling.

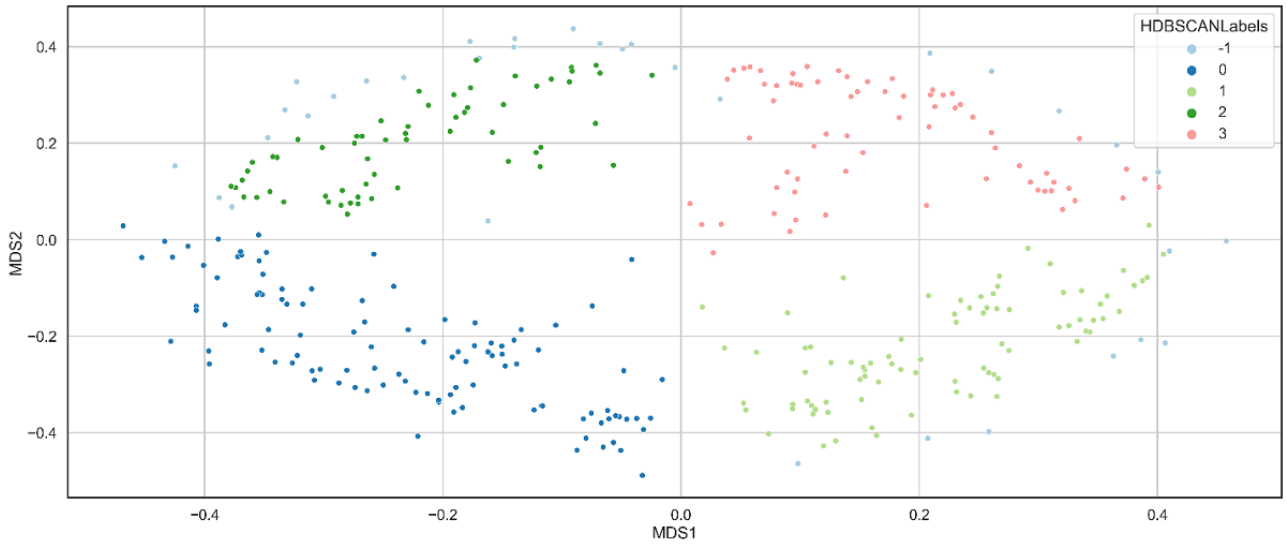


Figure 2. Condensed tree resulting from the hierarchical density-based spatial clustering of applications with noise algorithm performed on the data. Note: similar to a classical dendogram in a hierarchical clustering setting, the first yellow rectangle represents the entire data, which is split into two parts (called “branches”) when we reduce the maximum distance allowed between points within each branch (λ value = 1 / distance). Each rectangle represents a subpart of the data after a split and with a size proportional to the number of data points in the subpart. The entire data splits into cluster 0 and the green rectangle, which further splits into cluster 1 and a turquoise rectangle, when we reduce the distance allowed. The 4 detected clusters (signified by a circle and their number) are the branches that persist the most (do not split further, according to various rules of the algorithm) when the imposed maximum distance between points decreases while keeping a minimum size. The persistence is proportional to the length of the rectangles across the vertical axis. The tree can be interpreted as a probability distribution function upside down, with each cluster being a peak in the distribution.

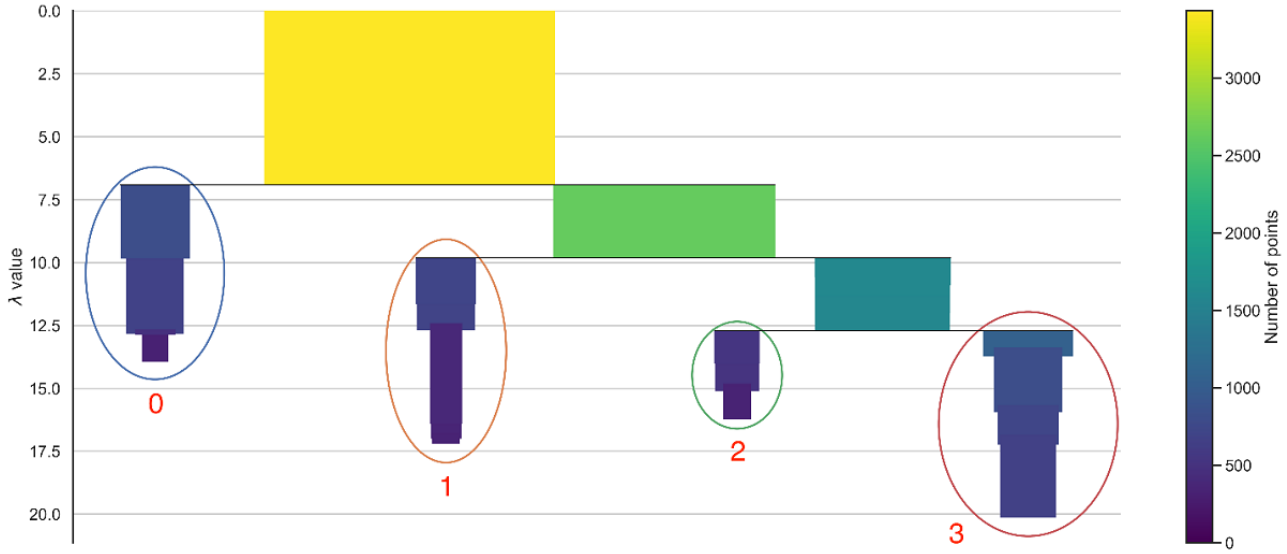


Table 1. Descriptive statistics of clusters.

Statistics	All data	Outliers	Cluster 0 "Complex high-cost high-need"	Cluster 1 "Slightly complex with inexpensive low-severity PCGs[a]"	Cluster 2 "Oldest at high risk"	Cluster 3 "Patients with 1 costly disease"	No PCGs	Hypertension "Only hypertension"	Mental health "Only mental diseases"
Patients, n (%)	18,732 (100.0)	321 (1.7)	817 (4.4)	709 (3.8)	531 (2.8)	1056 (5.6)	12,720 (67.9)	1813 (9.7)	765 (4.1)
Age (years), mean (SD)	65.0 (10.6)	66.3 (10.8)	66.3 (10.6)	67.8 (10.2)	69.4 (10.9)	68.1 (11.2)	64.0 (10.4)	67.6 (9.7)	63.2 (10.9)
Sex, n (%)									
Men	8626 (46)	130 (40)	325 (40)	205 (29)	279 (53)	536 (51)	5772 (45)	1158 (64)	221 (29)
Women	10,106 (54)	191 (60)	492 (60)	504 (71)	252 (47)	520 (49)	6948 (55)	655 (36)	544 (71)
Deductible (CHF; US $), mean	794 (852)	511 (548)	448 (481)	535 (574)	524 (562)	562 (603)	908 (974)	612 (657)	558 (599)
Model with gatekeeper[b]	0.5	0.4	0.4	0.4	0.4	0.4	0.5	0.5	0.5
Number of PCGs, mean	0.4	1.2	2.1	1.7	1.3	1.1	0.0	1.0	1.0
Multimorbid (yes)[b]	0.1	0.1	0.8	0.6	0.3	0.1	0.0	0.0	0.0
Ambulatory costs (CHF; US $), mean	5395 (5789)	7967 (8549)	11,731 (12,589)	7477 (8024)	9728 (10,439)	10,362 (11,120)	4074 (4372)	5462 (5861)	7571 (8125)
Inpatient costs (CHF; US $), mean	1419 (1523)	2134 (2290)	3109 (3336)	1811 (1943)	2749 (2950)	1575 (1690)	1199 (1287)	1372 (1472)	1585 (1701)
Costs of medications (CHF; US $), mean	1563 (1677)	2683 (2879)	4073 (4371)	2221 (2383)	3587 (3849)	4450 (4775)	965 (1036)	1732 (1859)	1961 (2104)
Total cost (CHF; US $), mean	8929 (9582)	13,684 (14,684)	19,950 (21,409)	12,440 (13,349)	17,057 (18,304)	17,312 (18,578)	6611 (7094)	9439 (10,129)	12,025 (12,904)
Number of days in the hospital, mean	2.6	4.3	6.6	3.6	5.6	3.4	2.0	2.4	3.5
Number of hospitalizations in a year, mean	0.2	0.4	0.5	0.3	0.4	0.3	0.2	0.3	0.3
Total number of consultations, mean	11.9	16.0	20.2	17.0	17.5	16.1	9.9	12.7	18.5
Number of consultations with generalist, mean	7.2	10.0	11.6	9.8	11.3	9.4	6.0	8.3	9.5
PCG groups in the cluster	All 34 PCGs	Mostly Pain	Mental + hypertension + pain + asthma (COPD[c])	Thyroid + hypertension + glaucoma + mix of others	Asthma + Parkinson + cardiac diseases + pain	Cancer + diabetes + inflammatory + immune + other mental + glaucoma + HIV	N/A[d]	Hypertension	Mental diseases

Statistics	All data	Outliers	Cluster 0 “Complex high-cost high-need”	Cluster 1 “Slightly complex with inexpensive low-severity PCGs[a]”	Cluster 2 “Oldest at high risk”	Cluster 3 “Patients with 1 costly disease”	No PCGs	Hypertension “Only hypertension”	Mental health “Only mental diseases”
Description of the clusters based on overall descriptive statistics	N/A	Average age, slightly fewer male patients, higher hospital costs and hospital stays	Average age, slightly fewer male patients, lowest deductibles, highest amount of PCGs and multimorbidity, highest health care use and costs (except for costs of medications)	Slightly older, more female patients, relatively low deductibles, high amount of PCGs (1.7) and multimorbidity (but less than cluster 0), relatively low health care use and costs	Oldest, relatively low deductibles, some complexity (more than 1 PCGs on average), very high use of doctor visits (especially generalist), many hospitalizations and high inpatient costs	Relatively old, on average 1 PCG, highest cost of medicaments, and high ambulatory costs, relatively low hospitalizations and doctor visits	Young, highest deductibles, low health care use and costs	Slightly older, more male patients, relatively low health care use and costs	Youngest, more female patients, relatively low deductibles, low health care use and costs (but higher than for hypertension group), a lot of visits to doctors

[a]PCG: pharmacy-based cost group.

[b]Ratios rounded off to one decimal place.

[c]COPD: chronic obstructive pulmonary disease.

[d]N/A: not applicable.

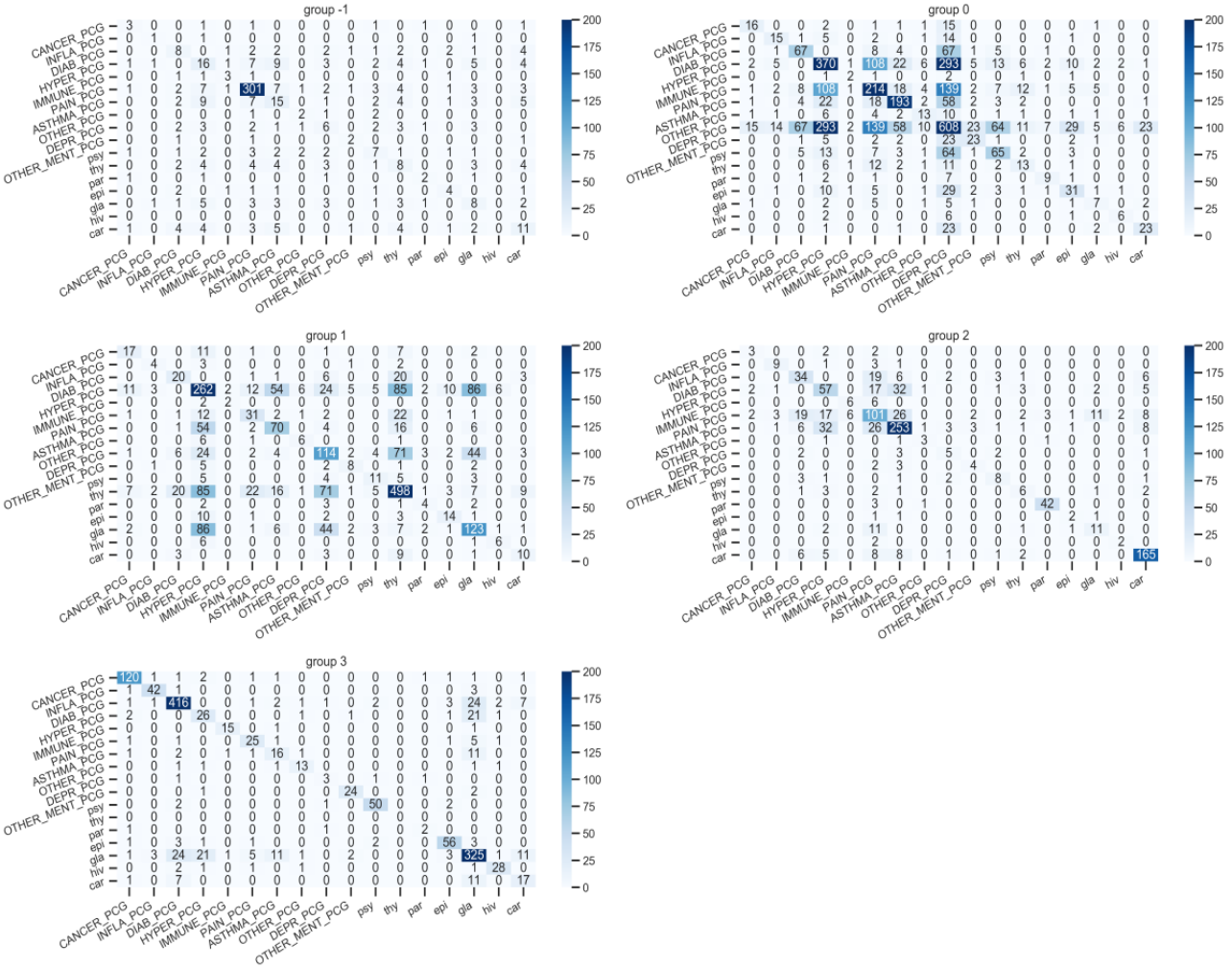


Figure 3. Joint distributions of PCGs within the 4 clusters (group 0-3) and outliers (group –1). PCG: pharmacy-based cost group.

Discussion

Our study shows that performing cluster analysis to explore patient multimorbidity and complexity is feasible. We demonstrated that individuals with single PCGs of mental diseases or hypertension, individuals with multiple PCGs, or individuals with a single high-cost PCG have different health care use patterns and represent different complexity groups.

Earlier studies focusing on chronic conditions identified from electronic health records evidenced the existence of systematic associations between chronic diseases, whereby chronic diseases, often from dissimilar disease categories, coappeared within a multimorbidity pattern or cluster [24-26]. Importantly, though, these studies showed that the complexity of multimorbidity patterns in terms of diseases and associated drug use increased with age, which holds true for both genders. Moreover, in line with our findings, multiple earlier studies used cluster analysis for identifying clinically homogenous multimorbidity patterns in the population, where clusters were composed of diagnosis-related groups [16,27-30]. However, these studies used measures of multimorbidity and comorbidity or clinical diagnosis data rather than PCGs from claims data. This makes direct comparison of results challenging, due to the differences in methodologies and level of diagnosis details. A recent systematic review confirmed that analytical methods used to identify patient profiles with multimorbid conditions are heterogeneous (including factor analysis, multiple correspondence analysis, hierarchical clustering, and three-step unified-clustering method), which may explain the variation in the multimorbidity patterns reported in various studies [31]. Despite those differences, the observed most prevalent clusters or groups are similar across studies and included hypertensive or metabolic diseases [28,29] and mental and behavioral diseases [16]. The greater prevalence of and similarities in metabolic and mental clusters were confirmed by a systematic review of multimorbidity patterns, whereby these clusters were identified in 9 and 10 of 14 reviewed articles, respectively [32]. One study compared multimorbidity patterns between populations of two European countries (Spain and the Netherlands) and found that, indeed, the highest similarities were observed in the cardio metabolic cluster, even though the populations are likely to differ across countries [26].

The existing literature on the use of cluster analysis to identify homogenous segments based on health care use and expenditures is limited [33-37]. Specifically, the study by Nnoaham and Cann [33] identified segments (or clusters), similar to ours, based on health care use (expressed by visits to the physicians, medications, and admissions) and complexity (expressed by long-term conditions). Other studies used cluster analysis to identify groups with high expenditures and deduced that, despite

having a lot of heterogeneity, the high expenditures cluster typically exhibited fair or poor health with more medical conditions or comorbidities [34,35]. These findings confirm ours; they nevertheless need to be interpreted with caution due to differences in methodologies, age of the population, and level of details available for background individual characteristics and diagnoses. There is evidence that cluster analysis may provide more information to decision makers than a list of possible statistically significant variables or a list of individuals who are the highest users [35].

To our knowledge, this is the first study using cluster analysis to explore patients' multimorbidity and complexity, reflected by the mix of PCGs and health care use patterns. In addition, it benefits from the richness of health care use data, a large sample size, and advanced clustering methods. However, the study has certain limitations. The first limitation stems from the process of multiple parameters configuration, which increases complexity while not allowing results validation. Thus, the cluster interpretation has to rely on metrics from the algorithms, descriptive statistics, and clinical relevance. Second, as the data were lacking clinical information, we only relied on PCGs mapping, which may give an incomplete picture of drug data [9,11,12].

Our study shows that PCG-based cluster analysis of health care use claims data allows diverting from an approach of simple comorbidity counts and can identify the population profiles with increased health care use and costs. Such results may provide insightful information for policy making, care planning, and care delivery to facilitate the transformation from procedures and guidelines focusing on a single disease toward development of integrated and better coordinated care.

Acknowledgments

This work was supported by the Swiss National Science Foundation within the Smarter health care—National Research Programme (NRP 74) and grant 407440_183447.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Lists of pharmacy-based cost groups used to identify chronic diseases in insurance claims data.
[DOCX File , 15 KB - medinform_v10i4e34274_app1.docx]

References

1. Prince MJ, Wu F, Guo Y, Gutierrez Robledo LM, O'Donnell M, Sullivan R, et al. The burden of disease in older people and implications for health policy and practice. Lancet 2015 Feb 07;385(9967):549-562. [doi: 10.1016/S0140-6736(14)61347-7] [Medline: 25468153]
2. Bachmann N, Burla L, Kohler D. La santé en Suisse – Le point sur les maladies chroniques: Rapport national sur la santé 2015. OBSAN. Berne: Hogrefe Verlag; 2015. URL: https://www.obsan.admin.ch/sites/default/files/2021-08/rapportsante_2015_f_0.pdf [accessed 2022-03-16]
3. Smith SM, Wallace E, O'Dowd T, Fortin M. Interventions for improving outcomes in patients with multimorbidity in primary care and community settings. Cochrane Database Syst Rev 2021 Jan 15;1:CD006560 [FREE Full text] [doi: 10.1002/14651858.CD006560.pub4] [Medline: 33448337]
4. Smith SM, Wallace E, O'Dowd T, Fortin M. Interventions for improving outcomes in patients with multimorbidity in primary care and community settings. Cochrane Database Syst Rev 2016 Mar 14;3:CD006560 [FREE Full text] [doi: 10.1002/14651858.CD006560.pub3] [Medline: 26976529]
5. Souza DLB, Oliveras-Fabregas A, Minobes-Molina E, de Camargo Cancela M, Galbany-Estragués P, Jerez-Roig J. Trends of multimorbidity in 15 European countries: a population-based study in community-dwelling adults aged 50 and over. BMC Public Health 2021 Jan 07;21(1):76 [FREE Full text] [doi: 10.1186/s12889-020-10084-x] [Medline: 33413239]
6. Pefoyo AJK, Bronskill SE, Gruneir A, Calzavara A, Thavorn K, Petrosyan Y, et al. The increasing burden and complexity of multimorbidity. BMC Public Health 2015 Apr 23;15:415 [FREE Full text] [doi: 10.1186/s12889-015-1733-2] [Medline: 25903064]
7. Amelung V, Stein V, Suter E, Goodwin N, Nolte E, Balicer R, editors. Handbook Integrated Care. Cham: Springer; 2017.
8. Sharabiani MTA, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Med Care 2012 Dec;50(12):1109-1118. [doi: 10.1097/MLR.0b013e31825f64d0] [Medline: 22929993]
9. Huber CA, Szucs TD, Rapold R, Reich O. Identifying patients with chronic conditions using pharmacy data in Switzerland: an updated mapping approach to the classification of medications. BMC Public Health 2013 Oct 30;13:1030 [FREE Full text] [doi: 10.1186/1471-2458-13-1030] [Medline: 24172142]
10. Huber CA, Schneeweiss S, Signorell A, Reich O. Improved prediction of medical expenditures and health care utilization using an updated chronic disease score and claims data. J Clin Epidemiol 2013 Oct;66(10):1118-1127. [doi: 10.1016/j.jclinepi.2013.04.011] [Medline: 23845184]

11. Lamers LM, van Vliet RCJA. The Pharmacy-based Cost Group model: validating and adjusting the classification of medications for chronic conditions to the Dutch situation. Health Policy 2004 Apr;68(1):113-121. [doi: 10.1016/j.healthpol.2003.09.001] [Medline: 15033558]
12. Chini F, Pezzotti P, Orzella L, Borgia P, Guasticchi G. Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. BMC Public Health 2011 Sep 05;11:688 [FREE Full text] [doi: 10.1186/1471-2458-11-688] [Medline: 21892946]
13. Whitson HE, Johnson KS, Sloane R, Cigolle CT, Pieper CF, Landerman L, et al. Identifying patterns of multimorbidity in older Americans: application of latent class analysis. J Am Geriatr Soc 2016 Aug;64(8):1668-1673 [FREE Full text] [doi: 10.1111/jgs.14201] [Medline: 27309908]
14. Khalid S, Prieto-Alhambra D. Machine learning for feature selection and cluster analysis in drug utilisation research. Curr Epidemiol Rep 2019 Jul 27;6(3):364-372. [doi: 10.1007/s40471-019-00211-7]
15. Wartelle A, Mourad-Chehade F, Yalaoui F, Chrusciel J, Laplanche D, Sanchez S. Clustering of a health dataset using diagnosis co-occurrences. Appl Sci 2021 Mar 07;11(5):2373. [doi: 10.3390/app11052373]
16. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract 2018 Jul 03;19(1):108 [FREE Full text] [doi: 10.1186/s12875-018-0790-x] [Medline: 29969997]
17. Polynomics AG. Aktualisierung der PCG-Liste für den Schweizer Risikoausgleich. Studie im Auftrag des Bundesamts für Gesundheit BAG - Schlussbericht. 2019. URL: https://www.bag.admin.ch/dam/bag/en/dokumente/kuv-aufsicht/pus/risikoausgleich/corrigendun.pdf.download.pdf/Polynomics_Uni_Basel_Aktualisierung_PCG_Schlussbericht_2019-01-22.pdf [accessed 2020-03-09]
18. Breiman L. Random forests. Machine Learning 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
19. Breiman L, Cutler A. Random Forests Manual v4.0. UC Berkeley. 2003. URL: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf [accessed 2020-10-30]
20. Shi T, Seligson D, Belldegrun AS, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 2005 Apr;18(4):547-557. [doi: 10.1038/modpathol.3800322] [Medline: 15529185]
21. Allen E, Horvath S, Tong F, Kraft P, Spiteri E, Riggs AD, et al. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. Proc Natl Acad Sci U S A 2003 Aug 19;100(17):9940-9945 [FREE Full text] [doi: 10.1073/pnas.1737401100] [Medline: 12909712]
22. Kruskal JB, Wish M. Multidimensional scaling. In: Multidimensional Scaling. Quantitative Applications in the Social Sciences, ed. I. Thousand Oaks, CA: SAGE Publications; 1978.
23. McInnes L, Healy J, Astels S. hdbscan: hierarchical density based clustering. J Open Source Software 2017 Mar;2(11):205. [doi: 10.21105/joss.00205]
24. Ioakeim-Skoufa I, Poblador-Plou B, Carmona-Pírez J, Díez-Manglano J, Navickas R, Gimeno-Feliu LA, et al. Multimorbidity patterns in the general population: results from the EpiChron cohort study. Int J Environ Res Public Health 2020 Jun 14;17(12):4242 [FREE Full text] [doi: 10.3390/ijerph17124242] [Medline: 32545876]
25. Mucherino S, Gimeno-Miguel A, Carmona-Pirez J, Gonzalez-Rubio F, Ioakeim-Skoufa I, Moreno-Juste A, et al. Changes in multimorbidity and polypharmacy patterns in young and adult population over a 4-year period: a 2011-2015 comparison using real-world data. Int J Environ Res Public Health 2021 Apr 21;18(9):4422 [FREE Full text] [doi: 10.3390/ijerph18094422] [Medline: 33919351]
26. Poblador-Plou B, van den Akker M, Vos R, Calderón-Larrañaga A, Metsemakers J, Prados-Torres A. Similar multimorbidity patterns in primary care patients from two European regions: results of a factor analysis. PLoS One 2014;9(6):e100375 [FREE Full text] [doi: 10.1371/journal.pone.0100375] [Medline: 24956475]
27. Déruaz-Luyet A, N'Goran AA, Senn N, Bodenmann P, Pasquier J, Widmer D, et al. Multimorbidity and patterns of chronic conditions in a primary care population in Switzerland: a cross-sectional study. BMJ Open 2017 Jul 02;7(6):e013664 [FREE Full text] [doi: 10.1136/bmjopen-2016-013664] [Medline: 28674127]
28. Marengoni A, Rizzuto D, Wang H, Winblad B, Fratiglioni L. Patterns of chronic multimorbidity in the elderly population. J Am Geriatr Soc 2009 Feb;57(2):225-230. [doi: 10.1111/j.1532-5415.2008.02109.x] [Medline: 19207138]
29. Guisado-Clavero M, Roso-Llorach A, López-Jimenez T, Pons-Vigués M, Foguet-Boreu Q, Muñoz MA, et al. Multimorbidity patterns in the elderly: a prospective cohort study with cluster analysis. BMC Geriatr 2018 Jan 16;18(1):16 [FREE Full text] [doi: 10.1186/s12877-018-0705-7] [Medline: 29338690]
30. Egan BM, Sutherland SE, Tilkemeier PL, Davis RA, Rutledge V, Sinopoli A. A cluster-based approach for integrating clinical management of Medicare beneficiaries with multiple chronic conditions. PLoS One 2019;14(6):e0217696 [FREE Full text] [doi: 10.1371/journal.pone.0217696] [Medline: 31216301]
31. Ng SK, Tawiah R, Sawyer M, Scuffham P. Patterns of multimorbid health conditions: a systematic review of analytical methods and comparison analysis. Int J Epidemiol 2018 Oct 01;47(5):1687-1704. [doi: 10.1093/ije/dyy134] [Medline: 30016472]
32. Prados-Torres A, Calderón-Larrañaga A, Hancco-Saavedra J, Poblador-Plou B, van den Akker M. Multimorbidity patterns: a systematic review. J Clin Epidemiol 2014 Mar;67(3):254-266. [doi: 10.1016/j.jclinepi.2013.09.021] [Medline: 24472295]

33. Nnoaham KE, Cann KF. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? BMC Public Health 2020 May 27;20(1):798 [FREE Full text] [doi: 10.1186/s12889-020-08930-z] [Medline: 32460753]
34. Powers BW, Yan J, Zhu J, Linn KA, Jain SH, Kowalski JL, et al. Subgroups of high-cost Medicare advantage patients: an observational study. J Gen Intern Med 2019 Feb;34(2):218-225 [FREE Full text] [doi: 10.1007/s11606-018-4759-1] [Medline: 30511290]
35. Agterberg J, Zhong F, Crabb R, Rosenberg M. Cluster analysis application to identify groups of individuals with high health expenditures. Health Serv Outcomes Res Method 2020 Aug 01;20(2-3):140-182. [doi: 10.1007/s10742-020-00214-8]
36. Copeland LA, Zeber JE, Wang C, Parchman ML, Lawrence VA, Valenstein M, et al. Patterns of primary care and mortality among patients with schizophrenia or diabetes: a cluster analysis approach to the retrospective study of healthcare utilization. BMC Health Serv Res 2009 Jul 26;9:127 [FREE Full text] [doi: 10.1186/1472-6963-9-127] [Medline: 19630997]
37. Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. Popul Health Metr 2016 Nov 25;14:44 [FREE Full text] [doi: 10.1186/s12963-016-0115-z] [Medline: 27906004]

Abbreviations

DT: decision tree
HDBSCAN: hierarchical density-based spatial clustering of applications with noise
MDS: multidimensional scaling
OOB: out-of-bag
PCG: pharmacy-based cost group
RF: random forest

Edited by C Lovis; submitted 14.10.21; peer-reviewed by W Zhang, I Ioakeim-Skoufa; comments to author 20.12.21; revised version received 04.02.22; accepted 06.02.22; published 04.04.22.

Please cite as:
Nicolet A, Assouline D, Le Pogam MA, Perraudin C, Bagnoud C, Wagner J, Marti J, Peytremann-Bridevaux I
Exploring Patient Multimorbidity and Complexity Using Health Insurance Claims Data: A Cluster Analysis Approach
JMIR Med Inform 2022;10(4):e34274
URL: https://medinform.jmir.org/2022/4/e34274
doi:10.2196/34274
PMID:35377334

Original Paper

Patient Recruitment System for Clinical Trials: Mixed Methods Study About Requirements at Ten University Hospitals

Kai Fitzer[1,2*], MSc; Renate Haeuslschmid[3*], Dr. rer nat.; Romina Blasini[4], MSc; Fatma Betül Altun[5], MSc; Christopher Hampf[1], MSc; Sherry Freiesleben[1], MSc; Philipp Macho[6], MSc; Hans-Ulrich Prokosch[7], Prof Dr; Christian Gulden[7], MSc

[1]Core Unit Data Integration Center, University Medicine Greifswald, Greifswald, Germany

[2]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

[3]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[4]Institute of Medical Informatics, University of Giessen, Giessen, Germany

[5]Medical Informatics Group, University Hospital Frankfurt, Frankfurt, Germany

[6]Medical Informatics, Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

[7]Medical Informatics, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

[*]these authors contributed equally

Corresponding Author:
Kai Fitzer, MSc
Core Unit Data Integration Center
University Medicine Greifswald
Walter-Rathenau-Str 48
Greifswald, 17487
Germany
Phone: 49 383486 ext 7555
Email: kai.fitzer@uni-greifswald.de

Abstract

Background: Clinical trials are the gold standard for advancing medical knowledge and improving patient outcomes. For their success, an appropriately sized cohort is required. However, patient recruitment remains one of the most challenging aspects of clinical trials. Information technology (IT) support systems—for instance, patient recruitment systems—may help overcome existing challenges and improve recruitment rates, when customized to the user needs and environment.

Objective: The goal of our study is to describe the status quo of patient recruitment processes and to identify user requirements for the development of a patient recruitment system.

Methods: We conducted a web-based survey with 56 participants as well as semistructured interviews with 33 participants from 10 German university hospitals.

Results: We here report the recruitment procedures and challenges of 10 university hospitals. The recruitment process was influenced by diverse factors such as the ward, use of software, and the study inclusion criteria. Overall, clinical staff seemed more involved in patient identification, while the research staff focused on screening tasks. Ad hoc and planned screenings were common. Identifying eligible patients was still associated with significant manual efforts. The recruitment staff used Microsoft Office suite because tailored software were not available. To implement such software, data from disparate sources will need to be made available. We discussed concrete technical challenges concerning patient recruitment systems, including requirements for features, data, infrastructure, and workflow integration, and we contributed to the support of developing a successful system.

Conclusions: Identifying eligible patients is still associated with significant manual efforts. To fully make use of the high potential of IT in patient recruitment, many technical and process challenges have to be solved first. We contribute and discuss concrete technical challenges for patient recruitment systems, including requirements for features, data, infrastructure, and workflow integration.

(JMIR Med Inform 2022;10(4):e28696) doi:10.2196/28696

KEYWORDS

patient recruitment system; clinical trial recruitment support system; recruitment; patient screening; requirements; user needs; clinical trial; interview; survey; electronic support; clinical information systems; eHealth

Introduction

Medical research requires the involvement of sufficiently sized and eligible patient cohorts. A shortage of participants may result in delays, reduced statistical validity, increased costs, or even the failure of costly trials [1]. Indeed, poor recruitment has been found to be the main reason for trial discontinuation [2,3]. Only 31% of the analyzed clinical trials were able to reach the targeted participant count within the time frame [4,5]. Williams et al [6] analyzed ended trials published on ClinicalTrials.gov and concluded that 57% of those were terminated owing to an insufficient rate of accrual. Those numbers point to a strong need for support, and software tools are a promising approach that may improve effectiveness and efficiency [7-12]. As noted by Beresniak et al [7], optimizing processes and using efficient IT support could reduce costs and allow for more clinical trials to be successfully completed with fewer resources. Software can support the recruitment process in various ways. In this work, we are particularly interested in tools that support the recruitment staff in identifying eligible patients by screening large amounts of routine data, including screening electronic health records (EHRs) for specific criteria. Those systems are commonly called patient recruitment systems (PRS), clinical trial recruitment support systems (CTRSS), or sometimes also called clinical decision–support systems [13].

There are numerous studies on approaches, prototypes, and tools to support patient recruitment. Cuggia et al [12] compared 28 PRS regarding their contributions, limitations, features, and efficacy. Köpcke et al [14] reviewed 79 PRS regarding their design and theorized on the approaches; for example, why they think specific design decisions were taken. A very recent review of Pung and Rienhoff [1] included 36 articles that evaluated recruitment-related electronic systems or described related workflows. A major caveat of these works is that neither their efficiency nor their design have been sufficiently evaluated in a real-world, clinical environment. These publications note improvements in recruitment in terms of effectiveness and efficiency; for example, increased recruitment numbers and time savings. However, only a few systems have been subjected to meaningful evaluation to date.

In their review, Köpcke et al [14] concluded that the utility of a PRS depends on the patient data available and the integration of the PRS into study and clinical workflows, but they also say that this has not been sufficiently explored. Cuggia et al [12] showed that the workflow, organization, and communication as well as users' perception and acceptance have not yet been sufficiently considered. Based on their review of the matter, they identified the physicians' limited time as well as their knowledge and awareness of trials to be some of the major obstacles.

Straube et al [15] identified limited human and technical resources and the high documentation effort as the most prevalent barriers. Schreiweis and Bergh [16] argue that “a detailed analysis of stakeholders’ requirements would help implementing better patient recruitment systems (PRS) in the future” and took a 2-fold approach to do so. They named some functional requirements for PRS, but no information on the context, the methods, or the participant count (N) was provided. Trinczek et al [17] interviewed 6 “domain experts” and identified a set of 23 tasks, with most tasks being related to patient identification. In a later study, Trinczek et al [18] showed that a large proportion of the work is manual and paper-based because the ability to search in the clinical databases is very restricted.

A PRS may be a solution to many of these issues. For an effective and accepted solution, we first need to thoroughly understand the users’ needs, current workflows, tasks, and barriers. Systems that are not well-embedded in the hospital work environment or those that do not answer direct needs are likely to be ineffective or even rejected by the users, which is why involving users in the design process is crucial for the success of these systems [19-21]. As listed above, few studies have evaluated the workflows and tasks, and they all have strong limitations; for example, they included few participants, provide little information about their methods and analysis, or only report on few aspects or requirements. We are not aware of a study that draws a complete picture and has holistically investigated the status quo in patient recruitment, including the recruitment workflows and tasks, as well as the users’ requirements and wishes regarding future PRSs. To extend the existing work, we applied a user-centered research approach and surveyed 56 prospective users and interviewed 33 potential ones—that is, patient recruitment staff—from 10 Medical Informatics in Research and Care in University Medicine (MIRACUM) university hospitals. We performed a qualitative analysis of the state-of-the-art recruitment workflows, procedures, issues, and existing technological support from the 10 sites. Furthermore, we established a collection of user-centric requirements for future patient recruitment systems. We completed this paper with a discussion of technical and functional requirements as well as how and where a PRS may be integrated in the clinical infrastructures and processes.

Methods

Methods Overview

We aimed to assess the status quo in patient recruitment to better support it with appropriate IT systems. We predominantly collected quantitative data from a web-based survey and qualitative data based on interviews. Both, the web-based survey and interviews were developed and carried out simultaneously and took place at 10 university hospitals that are part of the MIRACUM consortium.

Web-based Survey

Design

The survey contained different questions about the current, local recruitment workflows, and specifically about the screening tasks and timing as well as the communication of recruitment suggestions. Furthermore, we were interested in different patient recruitment tools: we asked about their attitude toward such systems, the expected usefulness, and specific requirements. The initial set of questions were brainstormed collaboratively within the team. After these questions were transferred to the web-based survey tool, a pretest was conducted allowing all team-members to try the survey and report any issues and feedback. In total, the survey consisted of 16 questions with multiple-choice, rating scale, and free-text answer formats, which were structured thematically on 6 different pages. Each page contained between 3 and 7 items, where all items were mandatory to answer, but contained a "not applicable/can't say" option. The web-based survey was generated using SoSci Survey and captured data anonymously between December 2018 and June 2019 [22].

Participants

To capture as many different workflows and perspectives as possible, we aimed to recruit staff members who (1) were involved in the patient recruitment process and (2) filled different positions across a broad spectrum of wards. The survey was sent out to the members of the MIRACUM consortium, who then redirected it to researchers and clinical staff at their site.

Analysis

The survey had a completion rate of 93%. The statistical analysis was anonymously conducted using the R (version 3.6, The R Foundation) [23]. Since 30 questions had to be answered using a rating scale ranging from 1 to 5, we used the sjPlot package to visualize the results [24]. All multiple-choice questions were visualized with simple bar plots using the plotly package [25]. We only analyzed complete answers, and answers pertaining to free-text questions on work experience, job title, age, and work experience were manually preprocessed and grouped into common categories before plotting.

Interviews

Design and Procedure

By means of semistructured interviews, we aimed to gather qualitative insights into the workflows, procedures, tasks and other relevant aspects of patient recruitment. We considered the works of DeMoor [10] and Trinczek [18] when designing 14 questions targeting (1) status quo in recruitment, (2) existing technological support, (3) perceived quality and problems, (4) and requirements for a PRS. On average, the interviews took about 45 minutes.

Participants

Overall, we collected data from 33 participants from all 10 hospitals (2-7 interviewees per site). Face-to-face and voice-recorded interviews were conducted with 12 participants who gave written consent. Furthermore, we collected answers in free written form from 21 participants with whom scheduling an in-person interview was not possible such as to reach a larger number of participants.

Analysis

Before transcribing the voice-recorded interviews, we anonymized all data. We then applied a content analysis approach, as suggested by Mayring [26].

Two authors then read and independently coded 3 randomly selected interviews into codebooks. In this process, codes were assigned for the respective answers to the questions. If a researcher gave a very similar answer to a question, the same code was used. Afterward, the authors compared the independently created codes and merged the codebooks. Owing to a high coding agreement of 95%, the two authors then proceeded to code the remaining 30 interviews independently (15 each). In the case of incomplete interviews, only the answered questions were considered and also coded, as they contained valuable insights. Unanswered questions were not considered in the evaluation.

Ethical Approval

This study was ethically approved by the ethics committee of the Friedrich-Alexander-University Erlangen-Nürnberg (approval number 412_18B).

Results

Results Overview

Below, we report the results obtained from the interviews and the web-based surveys. We illustrated (1) the procedures currently implemented at the participating hospitals as well as (2) the requirements for future patient recruitment tools. We received 56 complete responses of doctors (n=26, 46%), study coordinators (n=7, 13%), study nurses (n=4, 7%), medical documentalists (n=4, 7%), study assistants (n=2, 4%), scientific and technical staff (n=1 each, 2%), and others (n=4, 7%). Seven participants (13%) did not specify their role. Fourteen respondents were aged 25-34 years, 25 were aged 35-44 years, and 10 were aged 45-54 years. The average number of working experiences in the field of patient recruitment was 12. The interviews were conducted between December 2018 and June 2019. The number of participants by medical specialty and participating site is shown in Table 1.

Table 1. Participants by medical specialty and participating site.

Medical specialty or department	Participants, n										
	Dresden	Erlangen	Frankfurt	Freiburg	Gießen	Greifswald	Magdeburg	Mainz	Marburg	Mannheim	Total
Obstetrics and gynecology	2	1	0	0	0	0	0	1	0	0	4
Internal medicine	1	0	1	2	0	1	1	3	0	2	11
Surgery	1	1	0	0	0	0	0	0	0	0	2
Pediatrics	0	1	0	0	1	0	0	0	0	0	2
Urology	0	0	1	0	0	0	1	0	0	1	3
Anesthesiology	0	0	0	0	1	0	0	1	0	0	2
Psychiatry	0	0	0	0	1	0	0	0	0	0	1
Medical genetics	0	0	0	0	1	0	0	0	0	0	1
Neurology	0	0	0	0	0	1	0	0	1	0	2
Radiation oncology	0	0	0	0	0	0	1	0	0	0	1
Ophthalmology	0	0	0	0	0	0	1	1	1	0	3
Coordination Center for Clinical Studies	1	0	0	0	0	1	0	0	1	0	3
(Comprehensive) Cancer Center	1	0	0	0	0	0	0	0	0	0	1
Total	6	3	2	2	4	3	4	6	3	3	

Current Recruitment Procedures and Infrastructure

Communication of Recruiting Trials

Our analysis revealed that the first hurdle was to ensure that all the involved parties were aware of a recruiting trial and its accompanying criteria. This awareness was raised through various channels that were either specific to a department and topic or to certain roles and duties. Our interviewees mentioned that they primarily learned about new trials through regular meetings (n=15), such as the tumor board review, through staff from the same clinic and department (n=9), through staff from other clinics and departments (n=9), through sponsors and industry partners (n=11), through clinical partners (n=2), or through personal networks (n=2). Furthermore, our interviewees mentioned that they learned about new studies in the context of training and courses (n=7) as well as at events of associations, fairs, and congresses (n=4), emails (n=3), telephone or SMS (n=2), or printed mail (n=1). Four interviewees highlighted that whether one knows about and is aware of a trial during everyday work depends on the interest and motivation of the employee.

Recruitment Procedures and Difficulties

In this section, we break down and summarize all recruitment procedures.

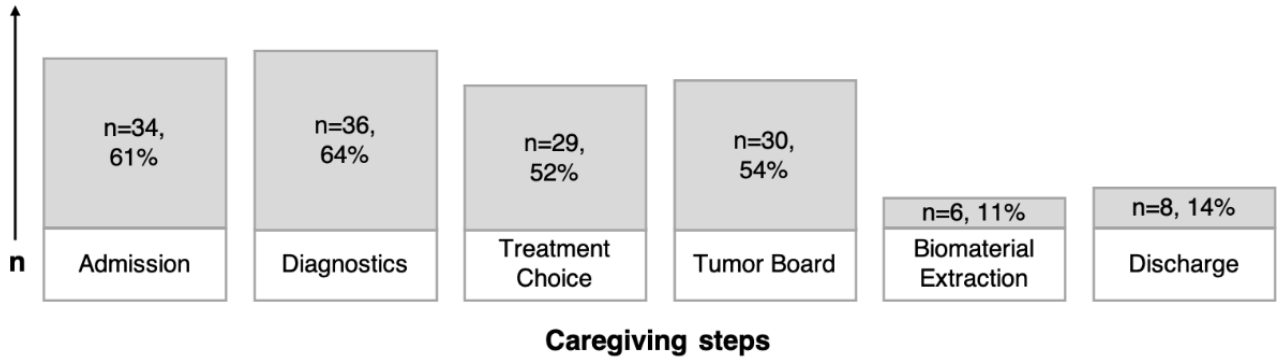
Procedures, Roles, and Tasks

Our interviews revealed that the clinical staff was mainly responsible to look for potential participants and forward suitable patient data for further screening (n=10). Research departments (n=2), coordinators (n=2), and auxiliary personnel (n=2) may also support the search; for example, by going through surgery schedules. When potential participants were identified, the research staff took over the detailed eligibility screening. Clinical investigators (n=8), study nurses (n=6), and assistants (n=3) may also be involved in this step. In contrast, our web-based survey exposed a different trend. According to our respondents, clinicians (n=44, 79%) and clinical investigators (n=45, 80%) were nearly equally involved in the identification of patients. Other staff members (n=31, 55%), study nurses and assistants (n=8, 14%) were also involved in identifying patients. Regarding the screening of patients, results from the interviews and web-based survey were more aligned. The survey showed that this task is assigned primarily to research staff; that is, clinical investigators (n=49, 88%) and study nurses and assistants (n=7, 13%). Furthermore, many of our survey respondents (n=22, 39%) stated that clinicians also take over certain screening duties.

Timing

In total, 23 interviewees mentioned that they followed a regular, cyclic, or daily recruitment procedure. In an ad-hoc manner, 4 interviewees screened newly moved patients to their ward, and 3 interviewees did not follow regular timing. The results of the web-based survey is presented in Figure 1, which shows the care-giving steps and the survey responses (as percentage values; multiple choices allowed). A patient's suitability for a trial was usually checked during admission, diagnosis, therapy choice, or during the tumor board, and less frequently checked when extracting or analyzing bio-materials, or upon discharge. A number of survey participants stated that they screened patients multiple times, either regularly (n=8, 14%) or without particular timing (n=24, 43%). In total, 19 respondents (34%) stated to screen a patient exactly once and 7 (13%) at every visit.

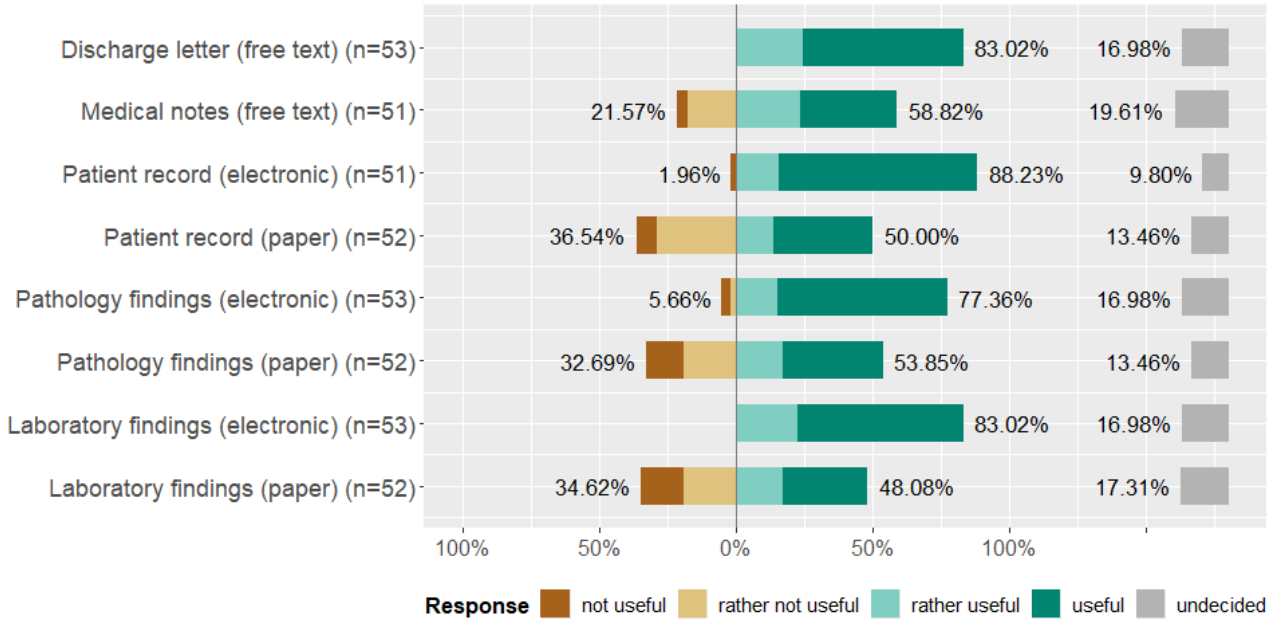
Figure 1. Patients are screened for eligibility at several stages during their hospital stay. The participants could select multiple stages.



Locations and Sources

Patients are recruited from their affiliated (n=17) and other departments (n=13), the emergency room (n=15), specific wards (n=9), and through the tumor board (n=12). Patients are recruited through external doctors (n=6) and partners (n=3), advertisements and press (n=5), and referred via other hospitals (n=2). Participants may be searched within (printed) ward-specific lists or schedules (n=4), patient files and medical reports (n=2 each). During the search, they first look at the department or clinic of the patients (n=7) as well as their diagnoses, procedures, laboratory samples and (existing) consents (n=1 each). Then, detailed screening in accordance with the eligibility criteria takes place. In the survey, our respondents rated different sources regarding their suitability for identifying potential participants from little useful (orange) to very useful (cyan). Figure 2 illustrates that electronic patient files, laboratory results, doctor's letters, and pathological findings are highly useful sources for finding participants. Paper-based or free-text data are regarded as slightly less useful.

Figure 2. Our participants rated a predefined set of data sources regarding their usefulness for identifying potential participants.



Difficulties

Overall, our interviewees judged the patient recruitment process as running very well (n=5), well (n=14), bad (n=5), or varying (n=4); for example, depending on the ward and staff. Many interviewees emphasized that sufficient direct communication between employees is essential to the recruitment procedure (n=18). They pointed out that staff-related problems such as staff turnover or shortfalls in motivation, communication, or support from external doctors (n=8) and logistic problems (n=6) are the most prevalent issues in the recruitment process. Furthermore, they mentioned that the data available in the hospital information system (HIS) were insufficient for screening (n=3) or that the eligibility criteria were too specific or complex to search via systems and databases (n=3).

In total, 46% of our interviewees stated that the identification of suitable patients hampers routine care. The screening procedure, consisting of searching for patients to checking all the eligibility criteria, was identified as the most time and labor-intensive step in recruitment (n=16). Informing participant candidates about the trial (n=10) and coordination tasks such as further diagnostics and data retrieval, questionnaires, and appointment management (n=9) were also considered time-consuming.

Infrastructure and Systems in Use

In total, 17 interviewees indicated that they already had systems deployed to support the recruitment process, and 13 said that they actually make use of them. Most of those systems were not dedicated to the recruitment process but rather tools developed for other duties and modified to fill the gap. To flag and document the recruited patients, various tools were used: Microsoft Office Tools (n=10), HIS and databases (n=12), SAP (n=7), patient files (n=5), papers (n=5), trial documents (n=5), and the tumor board (n=3). Four participants did not flag recruited patients. Our interviewees emphasized that the systems should provide a good overview (n=3), a good search and query opportunity (n=3), a good overall power (speed, data protection, and user and management function) (n=3). They complained about low data quality and, in particular, that data are insufficiently structured, outdated, and in need of further processing (n=2), which is why the resulting IT tools were not adequately functional.

Infrastructure Needs and Opportunities

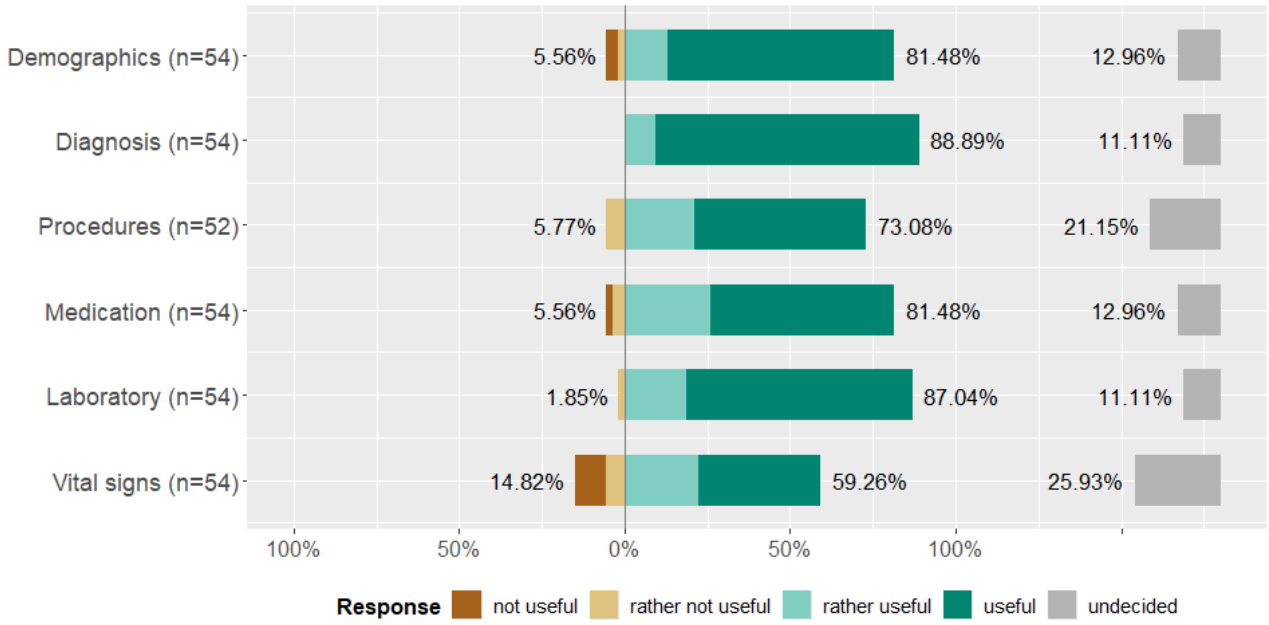
Patient Data

Our interviewees mentioned a broad spectrum of patient data that they would like to screen using IT support:

- Diagnoses (n=37): International Classification of Diseases–coded diagnoses (n=20), disease-specific values (n=7; eg, tumor values, heart values, scores, device data, and electrocardiographs), concomitant diseases (n=7), genetic information (n=2), and medical history (n=1)
- Demographic data (n=21): age (n=12), gender (n=6) and ethnicity, marital status, and place of residence (n=1 each)
- Treatment data (n=17): medication (n=6), therapy (n=6), Operation and Procedure Classification System–coded procedures (n=2), the date of surgery (n=2), and clinical findings (n=1)
- Laboratory data (n=12): blood count, heart failure markers, and histology; further laboratory values were indicated but not explicitly named
- Vital signs (n=8): patient's general condition (n=2), height (n=2), weight, implants, organ function, and study participation (n=1 each)

The survey results were overall in line with those of the interviews. Our survey respondents selected the diagnosis as the most important criteria, followed by laboratory data, demographics, medications, and procedures (summarized to treatment above). Vital signs also seemed useful, albeit with a lower priority. Figure 3 shows the respondents' estimated usefulness of the data rated on a 5-item scale.

Figure 3. Our participants rated a predefined set of data groups regarding their usefulness for patient identification.



Recruitment Suggestions

For recruitment suggestions, our interviewees wanted to be notified by a system (n=26) or to check suggestions by themselves (n=16). Three interviewees did not want to be notified. Notifications by email (n=9), popups (n=4), highlighting of the patient (n=3), SMS text messages or telephone (n=1), or in any way (n=8) were mentioned. Many interviewees desired a list of all patient suggestions, potentially integrated into the HIS (n=15). The survey results underline these desires: 83% of the respondents wanted a screening list with recruitment proposals and 81% wanted to receive email notifications.

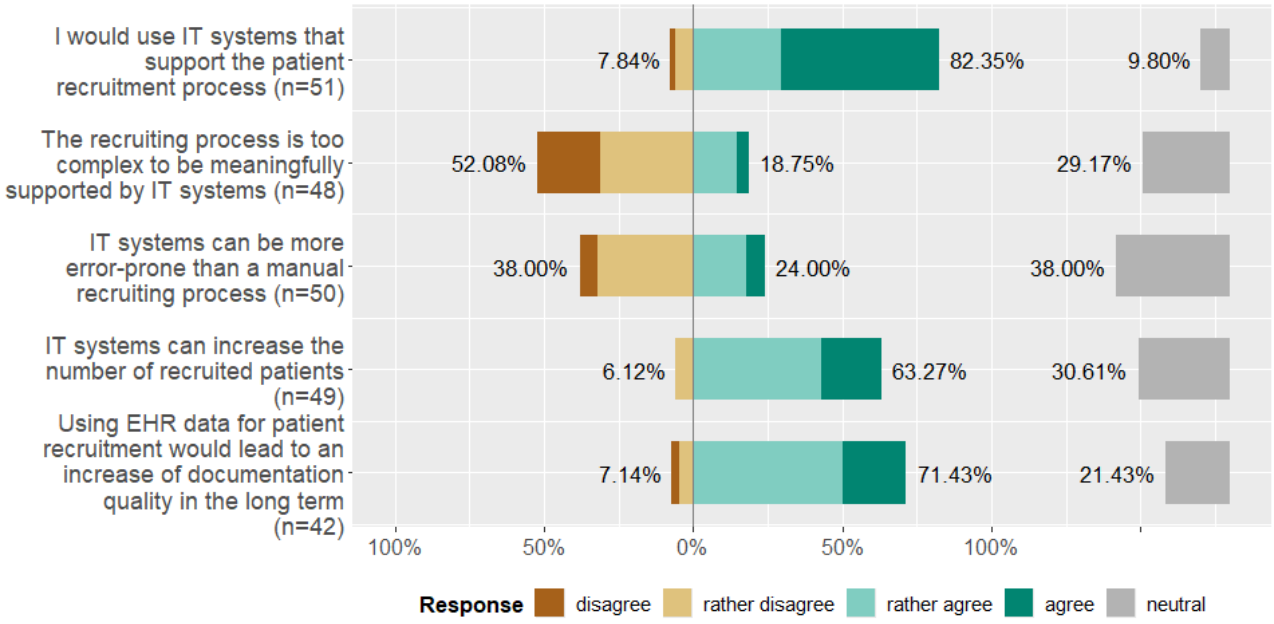
Wishes and Requirements

Our interviewees expressed other wishes; that is, they requested functional extensions and optimizations of existing clinical applications (n=18), more sophisticated searches for eligible patients (n=15), and an optimized tracking of included patients (n=3). They especially desired improvements regarding searching, usability, design, interoperability of research and clinical systems (n=9 overall), popups for requesting input of additional research data (n=6), and decision support (n=3).

Opinions and Expectations

As part of the web-based survey, we were interested in our participants' expectations concerning the capabilities of a PRS. Figure 4 illustrates the participants' selection of predefined multiple-choice answers in percentages, split in accordance with the 5-item scale. In general, our participants would want to use such a tool, and they assume that this type of system could be capable of supporting complex recruitment processes. They also expect that this tool would increase the number of recruited patients as well as the documentation quality, while potentially lowering errors.

Figure 4. Our participants' expectations of a patient recruitment system. EHR: electronic health record, IT: information technology.



Requirements for Patient Recruitment Systems

In the following section, we relate our findings to those reported in the literature and discuss functional and technical requirements toward future PRS.

Functional Requirements

We explicitly asked our interview participants to name their expectations and requirements toward a PRS. Below, we summarize the most relevant ones.

Overview of Patients

Almost all of our participants wanted an overview of potentially eligible patients for their study. Namely, one should be able to mark participants, make notes, and track the recruitment status. In addition, one should be able to manually add or remove patients from the recruitment list. Some participants specifically requested that patient summaries be integrated into existing systems.

Overview of Trials

Participants expressed that the PRS should include a module to manage studies and inclusion criteria, and that a link to the ClinicalTrials.gov study registry would also be useful. Many participants would also like to see a registry of all ongoing studies at the clinic. In some cases, cross-site recruitment support with other hospitals was desired.

Notifications

Furthermore, our participants wanted to instantly be notified when an eligible patient was found. Those notifications should also be manageable by the user to fit individual preferences as well as the workflow and roles [12,14].

Search

Additionally, our participants mentioned that a PRS absolutely needs to offer sophisticated search options; for example, for feasibility tests.

User Management and Interface

Our participants mentioned that the PRS must contain a sophisticated rights concept to account for the various roles in the study and at the clinical center, while the PRS must be easy to use by means of a clear and fast user interface.

Learn from Workarounds

Our interviewees mentioned that certain programs—for instance, Microsoft Office tools—are currently being used for patient recruitment, although these were not designed for this purpose. Future studies should therefore investigate how and for what these programs are used, so as to extract further functional requirements for the PRS.

Integration Into Workflows

Integrating PRS into clinical workflows can change various staff tasks and roles. Currently, many processes are carried out manually and are paper-based. For example, research staff might sift through recruitment proposals instead of manually screening all patient records as they do now. Physicians would no longer have to search for patients during rounds. However, we still assume that they would still be busy with other recruitment tasks, such as further diagnostics or patient education. However,

according to Good Clinical Practice, the final screening task should always be performed by a physician. A PRS could only make recruiting recommendations. Before a PRS is ready for real-world use, further research should examine its usability, effectiveness, and workflow integration or modification. Another round of development may be required to adapt the PRS to the new workflows and to ensure its acceptance and effectiveness.

Integration Into Technical Infrastructures

Schreiweis et al [16] suggest that a PRS needs to be integrated with existing systems, especially to avoid additional documentation burden on the staff. According to Campbell et al [4], a lack of integration of systems is one of the reasons why many studies do not achieve the required recruitment numbers. We argue that a PRS must be integrated or at least connected to the existing technical infrastructure to make data accessible to the PRS in a timely manner. This would also have the advantage of reusing central user and rights management and would avoid further tool changes. Often, data required for recruitment are distributed across multiple systems. Therefore, a PRS must either be able to handle disparate data sources or the facility establish and advance data integration protocol to create a unified, hospital-wide research repository. Using data from standardized research repositories (such as i2b2 or the Observational Medical Outcomes Partnership Common Data Model) has the major advantage that the PRS can be reused at any site implementing such a repository.

Data Availability and Accessibility

Our participants identified various criteria and data types that are needed to search and screen for eligible patients. They also mentioned that sophisticated searches in digital documents and data repositories are rarely possible. This inaccessibility may have various reasons, including the following:

1. Data are not collected, meaning that the data needed to compare a patient with trial criteria are not consistently collected for every patient and thus not available at all,
2. Data are analog (paper-based) or a digital version is not available and thus not accessible to the systems,
3. Data are unstructured, as in medical letters and thus not easily processed and searched,
4. Data quality is insufficient; for example, incomplete data or with documentation errors, and is thus not reliable,
5. Interoperability or rights are limited, meaning that data are present but in an inaccessible system,
6. Users are not provided with the right tools; for example, because there simply is no system that allows for a sophisticated search or the system is too complex for the users.

Each of these reasons points to specific criteria that are needed for the success of future PRS: relevant data must be collected, digitized, structured, quality-checked, and made available in a system that respects data privacy regulations and that is not too complex for the user. This is in line the findings of Trinczek et al [18]. Doods et al [27] developed a comprehensive clinical trial data inventory. They reviewed which data types were available in 9 European hospitals. Their results clearly show that hospitals are far from having complete data available for PRS. Nevertheless, their generated data lists can serve as an agenda for what data needs to be addressed and with what priority.

Discussion

Principal Findings

Low recruitment is one of the major reasons why clinical trials fail. Many studies indicate that patient recruitment systems can increase recruitment effectiveness and efficiency. To ensure that PRSs are successfully integrated in clinical environments in the long term, an in-depth analysis of the system context and requirements is needed [12,14,16]. Our study aims at identifying many aspects needed for a successful PRS and confirms many findings of related works, but also extends them in various ways. Similar to Becker et al [28], we found that approximately half of our participants do not use any software, and that most of those who do adopt a system (eg, Microsoft Excel), adopt one that is not intended for patient recruitment. Successful recruitment highly depends on the staff, particularly their motivation and knowledge [12,28,29] and interpersonal communication. A PRS, which can identify and screen patients, could change particular duties of staff members and possibly affect their workflows and collaboration. A future PRS will also need to be as flexible as the recruitment workflows, especially regarding when and how it is used. Similar to the study of Trinczek et al [18], our results show that a tremendous amount of work is done manually and is paper-based. Our results also confirm that highly specific searches in the clinical data repositories are not possible or very limited. Instead, our participants rely on various, often paper-based sources, such as consultation schedules and medical reports, to find and screen eligible patients. Doods et al [27] reported that only a fragment of the data needed for clinical trial feasibility studies is readily available and accessible in European hospitals. Furthermore, we show which data our participants regard as most important for patient recruitment as well as from which sources they get those data. It should be noted, however, that according to Gulden et al [30] not all data elements can be meaningfully queried by IT systems; for example, pregnancy status or capacity to consent is rarely documented.

We also demonstrated which concrete requirements a PRS needs to fulfill to be successful. Overall, our results confirm and extend the list of requirements reported by Schreiweis et al [16]. In addition, we have discussed various functional and technical requirements and provided concrete recommendations for the design, development, and integration of future PRS. Prospective users should be involved in the design and development process to ensure that the system meets their needs and capabilities. Sophisticated user studies should furthermore assess the quality of the systems well as their effectiveness for patient recruitment.

Limitations and Methodological Implications

In total, there are 33 university hospitals in Germany. In this study, we recruited 56 participants from 10 sites, which indicates that our sample is not necessarily representative of the status quo in Germany. We could neither include all hospitals nor recruit participants from each ward, department, and clinic from the hospitals involved in this study. Furthermore, the numbers

of participants in the interviews and web-based questionnaires were not evenly distributed across the sites. The web-based survey was answered by approximately 2-3 persons per site. Few sites were able to recruit additional participants, which adds more weightage to the responses of those sites. Since procedures can vary between wards as they vary between hospitals, we do not consider this problematic. On the contrary, to obtain highly representative findings, it was important to us to recruit a large number of clinicians from as many different specialties as possible. Further, we did not enforce the semistructured interviews to be conducted in person, causing some participants to opt for answering the questions in written form. This resulted in short or missing responses in some cases. We did not exclude those participants from our analysis as they all presented valuable insights. As some of those insights were only mentioned by a single participant, an exclusion of incomplete responses would mean to a loss of valuable findings. In a qualitative analysis, obtaining the same number of codes for every participant and every question can generally not be assured, which implies that even if all responses were complete, they might not contain more findings. Thus, we were able to obtain insights and requirements from a larger group of participants and specialist areas.

Conclusions and Future Prospects

Problems in patient recruitment are common in clinical trials. There are various ambitions to overcome this issue by means of a patient recruitment system, which supports the identification of potential participants. However, those attempts are not based on a profound investigation of the status quo of recruitment, the workflows and environment in which a PRS would have to be embedded, which risks user acceptance and therefore the success of such a system. We present detailed findings on the recruitment workflows, tasks, and timing. Furthermore, we report on the momentary IT support and discuss functional and technical requirements for patient recruitment systems. We showed that identifying eligible patients is still associated with significant manual effort. To enable the use of a PRS, data from disparate sources will need to be made available. Lastly, we contribute and discuss concrete technical challenges for patient recruitment systems, including requirements for features, data, infrastructure, and workflow integration. Regarding the next step, we suggest that our findings should be translated into interface and interaction concepts, which may then serve as a basis for development. We argue that users need to be involved in both steps, concept design and system testing, to ensure the success of the PRS.

Acknowledgments

This work was conducted within the MIRACUM (Medical Informatics in Research and Care in University Medicine) consortium. MIRACUM is funded by the German Ministry for Education and Research (FKZ 01ZZ1801A-M). This work was performed in (partial) fulfillment of the requirements for obtaining the degree Dr. rer. biol. hum. from the Friedrich-Alexander-Universität Erlangen-Nürnberg.

Authors' Contributions

KF evaluated the questionnaires and drafted major sections of the manuscript, coordinated the weekly polls, incorporated the comments and took over all steps of the copyediting process. RH guided and contributed to the analysis and drafted major sections of the manuscript. RB evaluated the questionnaire data and presented them in graphical format. FBA assisted in the creation of the graphs and revised the manuscript. SF checked and revised all text as a native English speaker. CH was involved in the evaluation of the results and revised the manuscript critically. PM assisted in the development of the document in our weekly meetings, reviews, and minor adjustments to the wording. HUP coordinated the study. CG created the initial version of the web-based questionnaire and semistructured interviews, contributed to the manuscript, and coordinated the study together with HUP. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Pung J, Rienhoff O. Key components and IT assistance of participant management in clinical research: a scoping review. JAMIA Open 2020 Oct;3(3):449-458 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa041] [Medline: 33215078]
2. Kasenda B, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, et al. Prevalence, characteristics, and publication of discontinued randomized trials. JAMA 2014 Mar 12;311(10):1045-1051 [FREE Full text] [doi: 10.1001/jama.2014.1361] [Medline: 24618966]
3. Briel M, Olu KK, von Elm E, Kasenda B, Alturki R, Agarwal A, et al. A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. J Clin Epidemiol 2016 Dec;80:8-15 [FREE Full text] [doi: 10.1016/j.jclinepi.2016.07.016] [Medline: 27498376]
4. Campbell M, Snowdon C, Francis D, Elbourne D, McDonald A, 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]

5. McDonald A, 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 07;7:9 [FREE Full text] [doi: 10.1186/1745-6215-7-9] [Medline: 16603070]
6. Williams RJ, Tse T, DiPiazza K, Zarin DA. Terminated Trials in the ClinicalTrials.gov Results Database: Evaluation of Availability of Primary Outcome Data and Reasons for Termination. PLoS One 2015;10(5):e0127242 [FREE Full text] [doi: 10.1371/journal.pone.0127242] [Medline: 26011295]
7. 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 Research (EHR4CR) European Project. Contemp Clin Trials 2016 Jan;46:85-91 [FREE Full text] [doi: 10.1016/j.cct.2015.11.011] [Medline: 26600286]
8. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009;48(1):38-44 [FREE Full text] [Medline: 19151882]
9. 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]
10. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162-173 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]
11. Schreiweis B, Trinczek B, Köpcke F, Leusch T, Majeed RW, Wenk J, et al. Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. Int J Med Inform 2014 Nov;83(11):860-868 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.08.005] [Medline: 25189709]
12. 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 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.02.003] [Medline: 21459664]
13. Embi P, 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]
14. Köpcke F, Prokosch H. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014 Jul 01;16(7):e161 [FREE Full text] [doi: 10.2196/jmir.3446] [Medline: 24985568]
15. Straube C, Herschbach P, Combs SE. Which Obstacles Prevent Us from Recruiting into Clinical Trials: A Survey about the Environment for Clinical Studies at a German University Hospital in a Comprehensive Cancer Center. Front Oncol 2017;7:181 [FREE Full text] [doi: 10.3389/fonc.2017.00181] [Medline: 28894695]
16. Schreiweis B, Bergh B. Requirements for a patient recruitment system. Stud Health Technol Inform 2015;210:521-525 [FREE Full text] [doi: 10.3233/978-1-61499-512-8-521] [Medline: 25991202]
17. 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 [FREE Full text] [Medline: 23920898]
18. Trinczek B, Köpcke F, Leusch T, Majeed R, 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]
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. Ouadahi J. A qualitative analysis of factors associated with user acceptance and rejection of a new workplace information system in the public sector: a conceptual model. Can J Adm Sci 2008 Sep;25(3):201-213. [doi: 10.1002/cjas.65]
21. Lin WT, Shao BB. The relationship between user participation and system success: a simultaneous contingency approach. Inf Manag 2000 Sep;37(6):283-295. [doi: 10.1016/s0378-7206(99)00055-5]
22. Leiner D. SoSci Survey (Version 3.1.06) [Computer software]. 2019. URL: https://www.soscisurvey.de [accessed 2021-12-17] [WebCite Cache ID 78sB7xiya]
23. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. European Environment Agency. URL: https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006 [accessed 2021-12-17] [WebCite Cache ID 77Z0Dx2Wq]
24. Daniel Lüdecke. sjPlot - Data Visualization for Statistics in Social Science. RDocumentation. URL: https://www.rdocumentation.org/packages/sjPlot/versions/2.8.7 [accessed 2021-12-17]
25. Plotly R Open Source Graphing Library. Plotly Graphing Libraries. URL: https://plotly.com/r/ [accessed 2021-12-17]
26. Mayring P. Qualitative Inhaltsanalyse: Grundlagen und Techniken. Weinheim, Germany: Julius Beltz GmbH & Co. KG; Feb 02, 2015.
27. 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]

28. Becker L, Ganslandt T, Prokosch H, Newe A. Applied Practice and Possible Leverage Points for Information Technology Support for Patient Screening in Clinical Trials: Qualitative Study. JMIR Med Inform 2020 Jun 16;8(6):e15749 [FREE Full text] [doi: 10.2196/15749] [Medline: 32442156]
29. Csimma C, Swiontkowski MF. Large clinical trials in musculoskeletal trauma: are they possible? Lessons learned from the international study of the use of rhBMP-2 in open tibial fractures. J Bone Joint Surg Am 2005 Jan;87(1):218-222. [doi: 10.2106/JBJS.D.01938] [Medline: 15634835]
30. Gulden C, Landerer I, Nassirian A, Altun FB, Andrae J. Extraction and Prevalence of Structured Data Elements in Free-Text Clinical Trial Eligibility Criteria. Stud Health Technol Inform 2019;258:226-230. [Medline: 30942751]

Abbreviations

EHR: electronic health record
HIS: hospital information system
IT: information technology
MIRACUM: Medical Informatics in Research and Care in University Medicine
PRS: patient recruitment system

Edited by C Lovis; submitted 12.03.21; peer-reviewed by A Scherag, P Wicks; comments to author 19.04.21; revised version received 25.06.21; accepted 04.12.21; published 20.04.22.

Please cite as:
Fitzer K, Haeuslschmid R, Blasini R, Altun FB, Hampf C, Freiesleben S, Macho P, Prokosch HU, Gulden C
Patient Recruitment System for Clinical Trials: Mixed Methods Study About Requirements at Ten University Hospitals
JMIR Med Inform 2022;10(4):e28696
URL: https://medinform.jmir.org/2022/4/e28696
doi:10.2196/28696
PMID:35442203

Original Paper

A Traditional Chinese Medicine Syndrome Classification Model Based on Cross-Feature Generation by Convolution Neural Network: Model Development and Validation

Zonghai Huang[1*], MA; Jiaqing Miao[2*], PhD; Ju Chen[1*], MA; Yanmei Zhong[1], MA; Simin Yang[3], MD; Yiyi Ma[1], MA; Chuanbiao Wen[1], MA

[1]College of Medical Information Engineering, Chengdu University of Traditional Chinese Medicine, Chengdu, China

[2]School of Mathematics, Southwest Minzu University, Chengdu, China

[3]College of Acupuncture-Moxibustion and Tuina, Chengdu University of Traditional Chinese Medicine, Chengdu, China

[*]these authors contributed equally

Corresponding Author:
Chuanbiao Wen, MA
College of Medical Information Engineering
Chengdu University of Traditional Chinese Medicine
37 Shierqiao Road
Chengdu, 610075
China
Phone: 86 18980069966
Email: 228237222@qq.com

Abstract

Background: Nowadays, intelligent medicine is gaining widespread attention, and great progress has been made in Western medicine with the help of artificial intelligence to assist in decision making. Compared with Western medicine, traditional Chinese medicine (TCM) involves selecting the specific treatment method, prescription, and medication based on the dialectical results of each patient's symptoms. For this reason, the development of a TCM-assisted decision-making system has lagged. Treatment based on syndrome differentiation is the core of TCM treatment; TCM doctors can dialectically classify diseases according to patients' symptoms and optimize treatment in time. Therefore, the essence of a TCM-assisted decision-making system is a TCM intelligent, dialectical algorithm. Symptoms stored in electronic medical records are mostly associated with patients' diseases; however, symptoms of TCM are mostly subjectively identified. In general electronic medical records, there are many missing values. TCM medical records, in which symptoms tend to cause high-dimensional sparse data, reduce algorithm accuracy.

Objective: This study aims to construct an algorithm model compatible for the multidimensional, highly sparse, and multiclassification task of TCM syndrome differentiation, so that it can be effectively applied to the intelligent dialectic of different diseases.

Methods: The relevant terms in electronic medical records were standardized with respect to symptoms and evidence-based criteria of TCM. We structuralized case data based on the classification of different symptoms and physical signs according to the 4 diagnostic examinations in TCM diagnosis. A novel cross-feature generation by convolution neural network model performed evidence-based recommendations based on the input embedded, structured medical record data.

Results: The data set included 5273 real dysmenorrhea cases from the Sichuan TCM big data management platform and the Chinese literature database, which were embedded into 60 fields after being structured and standardized. The training set and test set were randomly constructed in a ratio of 3:1. For the classification of different syndrome types, compared with 6 traditional, intelligent dialectical models and 3 click-through-rate models, the new model showed a good generalization ability and good classification effect. The comprehensive accuracy rate reached 96.21%.

Conclusions: The main contribution of this study is the construction of a new intelligent dialectical model combining the characteristics of TCM by treating intelligent dialectics as a high-dimensional sparse vector classification task. Owing to the standardization of the input symptoms, all the common symptoms of TCM are covered, and the model can differentiate the symptoms with a variety of missing values. Therefore, with the continuous improvement of disease data sets, this model has the potential to be applied to the dialectical classification of different diseases in TCM.

(JMIR Med Inform 2022;10(4):e29290) doi:10.2196/29290

KEYWORDS

intelligent syndrome differentiation; cross-FGCNN; TCM

Introduction

According to the 2019 edition of *LatestGlobal Medical Summary* released by the World Health Organization, TCM is evaluated as complementary and alternative medicine that can effectively prevent and treat a variety of diseases [1]. TCM has been tested and refined through thousands of years of medical practice, exerting extensive influence in East Asia and even in the world [2]. In clinical practice, the effectiveness of TCM has been significant for chronic diseases such as chronic obstructive pulmonary disease [3] and diabetes [4] as well as for gynecological conditions such as infertility [5] and dysmenorrhea [6]. However, TCM, an empirical product, lacks objective evaluation indicators. The treatment process of TCM is more like a black box, which makes people doubt its reliability, but TCM does manifest advantages in clinical practice. Similarly, the result of neural network classification is also a black box to obtain the practice of the process. Therefore, an increasing number of Chinese researchers have begun to apply a neural network to explore the treatment rules of TCM to further prove its objective effectiveness [7-9].

Syndrome differentiation is an important classification task. The treatment of diseases in TCM is subject to certain law: principle-methods-formulas-medicinal. The principle refers to the guidance of TCM theory. More specifically, under the guidance of TCM theory, patients' syndromes can be differentiated according to their symptoms, preferred treatment method is identified, an appropriate prescription is selected, and medication is chosen. In addition, correlation of all 4 examinations is essential to TCM treatment. Overall, the symptoms can be obtained from 4 diagnostic methods: inspection, auscultation and olfaction, inquiry, and palpation. From the point of view of machine learning, TCM dialectics can be regarded as a complex model, whose input is the 4 diagnostic information aspects of the patient and output is the syndrome type. With the advancement of machine learning technology, researchers have devoted themselves to the construction of this model. As traditional machine learning methods, decision tree [10,11], K-nearest neighbor [12], Bayes [13,14], and support vector machine (SVM) [15,16] have been widely used in intelligent dialectical tasks. In existing reports, these algorithms show satisfactory classification performance under complete data sets. However, electronic medical record data often have a significant amount of missing data. Missing data in these 4 models is a difficult problem to solve. With the rise of deep learning, neural networks have been gradually applied to such tasks [17,18]. With the deepening of the model, the amount of imbalanced data for different syndrome types is becoming more and more prominent. Due to the existence of rare syndrome types and unbalanced data sets of model diseases in TCM dialectics, over-fitting problems may occur in deep neural networks (DNNs). With the deepening of research, some researchers have further improved the training of dialectical models from the aspect of training data preprocessing to obtain a better fitting degree. From the point of symptom preprocessing, the 4 diagnosis information aspects were divided into multiple dimensions according to different categories [19]. Starting from the syndrome type, some researchers divide the syndrome type into smaller syndrome type factors [20]. These optimization methods solve the problem of data normalization to a certain extent but require more data integrity. In short, although existing models can distinguish their corresponding data sets well, they have high requirements with respect to data. Sufficient and complete patient information is required. In the real world, there is bound to be many missing data in patient information acquisition. Therefore, a model that is closer to the real world and that can effectively distinguish high-dimensional sparse data is needed.

The input dimension in the click-through-rate (CTR) task is large and sparse. Factorization machines (FMs) [21,22] obtain the relationship between 2 features by performing the inner product of the weights of the 2 corresponding features and automatically carrying out feature engineering. However, FMs are limited to a second-order cross-product in feature selection, hindering automatic selection of high-dimensional features. To automatically extract higher-order feature combinations, deep FMs classify each bit feature of the input into a field [23] based on the original FM model and construct a DNN in parallel to obtain high-order nonlinear features [24]. It can learn both low-order and high-order features, but it can only learn 2-dimensional and 1 high-dimensional feature and its coverage is not strong. With the advent of deep & cross network (DCN) [25], multidimensional cross features can be learned at the same time by using a cross network instead of FM. The DNN part of deep FM and DCN pays more attention to the nonlinear high-dimensional features generated by global data, ignoring some local features. Feature generation by convolution neural network (FGCNN) [26] uses a convolution neural network to extract local features and combines the advantages of the original multilayer perceptron (MLP) for global feature extraction, which allows the high-dimensional features of the model to contain more information. The success of the CTR model in the binary classification of high-dimensional sparse data inspired us to construct an improved dialectical multiclassification model suitable for the high-dimensional sparse symptom data of TCM.

Dysmenorrhea refers to severe pain in the lower abdomen before or during menstruation, which greatly affects the work, study, and life of women [27,28]. According to the presence or absence of pathological pelvic diseases, dysmenorrhea can be divided into primary dysmenorrhea and secondary dysmenorrhea [29]. At present, the main treatment of dysmenorrhea is nonsteroidal anti-inflammatory drugs or oral contraceptives; however, these medicines exert adverse effects on metabolism and the digestive system [30]. TCM has proven to be associated with fewer adverse effects and to have a more remarkable curative effect

on dysmenorrhea [31,32]. It is considered a safe and reliable alternative therapy for dysmenorrhea. In this paper, dysmenorrhea data were divided into fields according to the diagnosis module of TCM. A cross-FGCNN model was constructed, in which linear cross features were obtained by cross-layer and nonlinear high-dimensional features were generated by FGCNN. The contributions of this paper are as follows:

1) According to the thinking system of TCM diagnosis, a filed segmentation suitable for TCM dialectics was constructed so that the model can better fit different diseases.

2) Because of the large dimension and high sparsity of TCM symptom data, we used cross-layer to obtain multidimensional linear cross features and used FGCNN to obtain nonlinear high-dimensional features, including local and global features. As many features as possible were obtained from sparse data.

3) Training data and test data consisted of 4000 real dysmenorrhea clinical cases from Sichuan TCM big data management platform and 1273 dysmenorrhea cases from the Chinese literature database, so diversity of medical record data source was ensured. Two professional TCM doctors verified the data according to the relevant standards of TCM to ensure reliability of the data. To ensure the objectivity of this study, 6 traditional, intelligent dialectical models and 3 CTR models were selected and compared with our model in terms of accuracy, F1 score, confusion matrix, and log-loss.

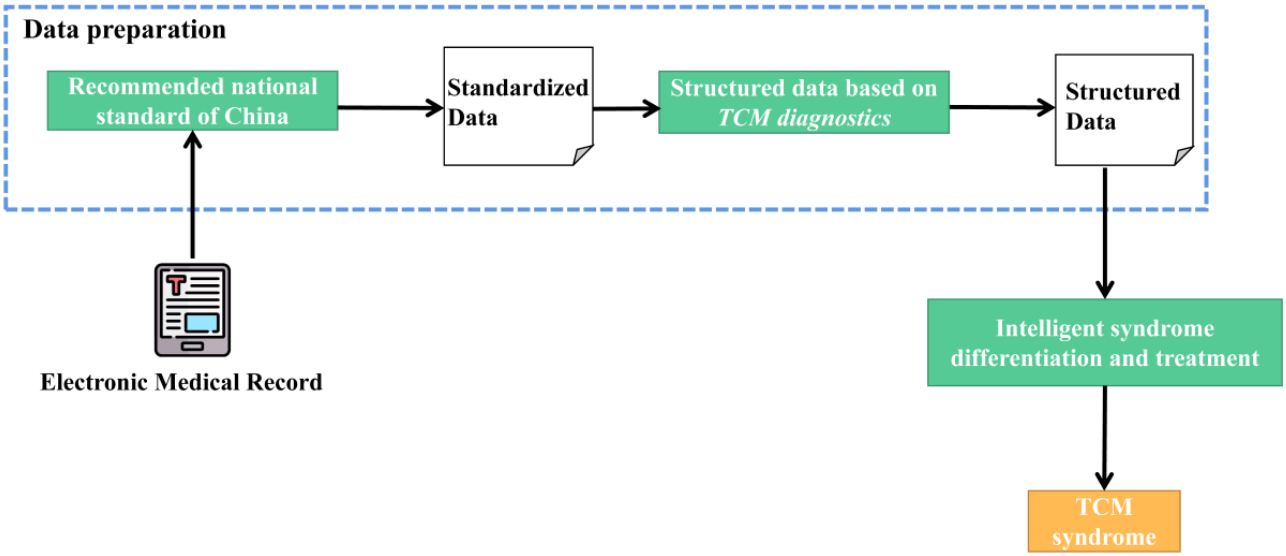
The structure of the rest of this paper is as follows: in the second section, we introduced our data acquisition, processing methods, and overall model structure; in the third section, we showed the experimental results; and in the last section, we put forward our conclusions.

Methods

Overview

Figure 1 shows the intelligent syndrome differentiation block diagram. Electronic medical records were first standardized. Then, standardized data were structured according to the classification of symptoms and physical signs in TCM diagnostics. The first 2 steps were the data preparation module. Finally, the prepared data were input to the intelligent dialectical model for TCM dialectical classification.

Figure 1. Intelligent syndrome differentiation block diagram. TCM: traditional Chinese medicine.



Data Preparation

A total of 4000 high-quality electronic medical records of dysmenorrhea were obtained from the Sichuan TCM big data management platform and 1273 cases of dysmenorrhea were obtained from the Chinese literature database. Any data related to the study were retained, such as symptoms, syndromes, and disease names. Before TCM doctors conduct syndrome differentiation, they synthesize the characteristics of many symptoms [28]. Therefore, in this study, professional TCM practitioners standardized the syndrome type and symptom dimension according to “GB/T20348-2006 TCM basic theory terminology” and “GB/T16751.2-1997 TCM Clinical diagnosis and treatment terminology-Syndrome part” issued by the State Administration of TCM. All symptoms related to the dialectics of TCM were classified according to the *Diagnostics of TCM*. For example, “white tongue coating” and “yellow tongue coating” were classified into inspection of tongue coating color. Later, the label encoder was carried out according to different symptom attributes. The overall division is shown in Table 1. Each symptom or physical sign represented an input dimension, so the overall input dimension was 60 dimensions. All symptoms or physical signs could not occur at the same time, so there must be a missing value and the missing value was −1.

Table 1. Classification of symptoms.

Diagnosis	Elements
Inspection (30 symptoms and physical signs in total)	• Expression • Complexion • Physique • Posture • Head • Face • Nose • Eye • Ear • Mouth • Tooth • Neck • Chest • Abdomen • Lumbar • Exterior genitalia • Anus • Skin • Phlegm • Saliva • Vomitus • Excrement • Urinating • Index fingers' superficial venules • Tongue nature • Tongue shape • Tongue color • Tongue coating nature • Tongue coating color • Hypoglossal vessels
Listening and smelling (6 symptoms and physical signs in total)	• Voice • Breathing sound • Snoring • Coughing sound • Belching • Tone
Inquiry (22 symptoms and physical signs in total)	• Cold and heat • Sweating • Pain site • Nature of pain • Head discomfort • Physical discomfort • Limb discomfort • Ear discomfort • Eye discomfort • Sleep • Diet • Thirst • Abnormal defecation • Abnormal urine • Menstrual period • Menstrual color • Menstrual volume • Menstrual nature • Emotion • Family history • Vaccination history • Physiological abnormality
Pulse feeling and palpation (2 symptoms and physical signs in total)	• Pulse condition • Pressing feeling

According to our dysmenorrhea data, the main syndrome types could be summarized into 9 types: liver-kidney depletion, pattern of congealing cold with blood stasis, cold-dampness stagnation, liver constraint and dampness-heat, deficiency of qi and blood, qi stagnation and blood stasis, kidney deficiency and blood stasis, dampness-heat stasis obstruction, and yang deficiency and internal cold. Therefore, the output of our classification model only focused on these 9 syndrome types. The proportion of the 9 syndrome types is shown in Table 2, which shows that our data were unevenly distributed.

Through standardization and structured operations, electronic medical data could be transformed into structured data. These data were used as input to the intelligent dialectical model. Figure 2 shows the preprocessing results of a real electronic medical record.

Table 2. Proportion of syndrome types (N=5273).

Syndrome type	Total, n (%)
Liver-kidney depletion	514 (9.7)
Pattern of congealing cold with stasis	720 (13.6)
Cold-dampness stagnation	568 (10.7)
Liver constraint and dampness-heat	522 (9.8)
Deficiency of qi and blood	575 (10.9)
Qi stagnation and blood stasis	751 (14.2)
Kidney deficiency and blood stasis	543 (10.2)
Dampness-heat stasis obstruction	544 (10.3)
Yang deficiency and internal cold	536 (10.1)

Figure 2. An example of electronic medical record preprocessing. EMR: electronic medical record.

EMR Notes

A 26 years old female who suffered from repeated cold pain in the lower abdomen during menstruation. Her menstrual cycle is regular with light menstruation. During menstruation, there are blood clots and she always feels cold in her extremities. Moreover, her tongue is purple with white tongue coating, whose the pulse is deep and rough.

Named Entities

Section name	Element name	Element value
Inspection	Coating color	white
	Coating nature	-1
	...	
Listening and smelling	Tone	-1
	Belch	-1
	...	
Inquiry	Nature of pain	Cold pain
	Pain site	Abdomen
	...	
Pulse feeling and palpation	Pulse condition	Deep pulse
	Pressing feeling	-1

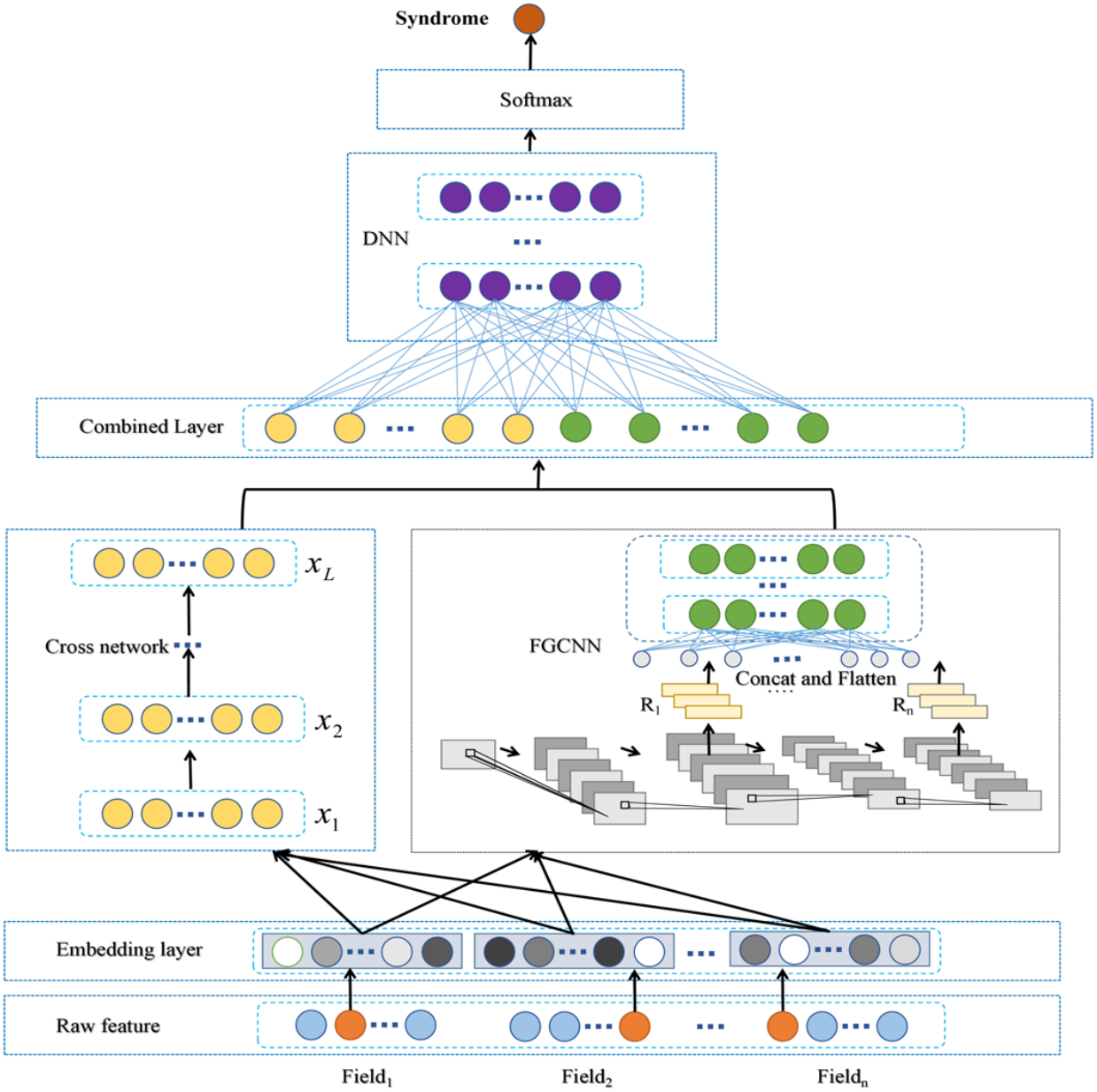
Intelligent Syndrome Differentiation Model: Cross-FGCNN

First, this section illustrates an overview of cross-FGCNN, which can automatically learn the feature interaction of high-dimensional and sparse symptom data to complete the intelligent dialectical task. Next, we describe in detail how to extract high-order combination features of low-dimensional representation from high-dimensional sparse vectors. Finally, data are classified.

Overall Structure of the Model

The whole model could be divided into 4 modules: data embedding module, cross linear feature extraction module, nonlinear feature extraction module, and classification module. The model read the symptom data found by label encoder and conducted a one-hot encoder according to each field. Then the embedding layer was used to map the high-dimensional sparse data to the low-dimensional dense features. The embedding data were used as shared input for the 2 parallel modules: the linear feature extraction module and the nonlinear feature extraction module. After the corresponding features were generated by the 2 modules, the 2 features were combined and input into the classification module, and finally the result syndrome type was obtained. The overall structure of the model is shown in Figure 3.

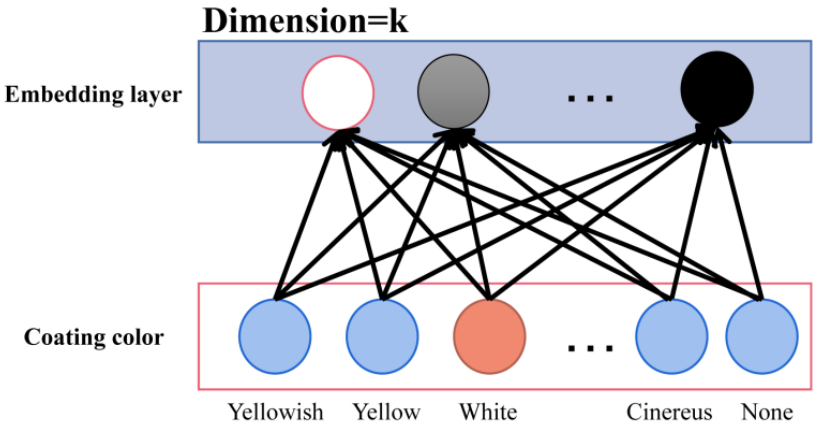
Figure 3. Model structure. DNN: deep neural network; FGCNN: feature generation by convolution neural network.



Data Embedding Module

The data embedding module of the model consisted of many structures as shown in Figure 4. According to the symptom and physical signs classification in *TCM diagnostics*, the obtained symptoms were mapped to different fields and one-hot coding was carried out. Embedding in the vector through dense embedding aimed to reduce the dimension of the embedding vector mapped from the field to the input model and ensured the density of the vector. For example, the physical sign, white tongue coating, was obtained from electronic cases and mapped to coating color field for encoding. The dimension of each field could be reduced to the specified dimension by the embedding operation, and the dimension of each field in the embedding layer was the same.

Figure 4. Coating color field of data embedding module.



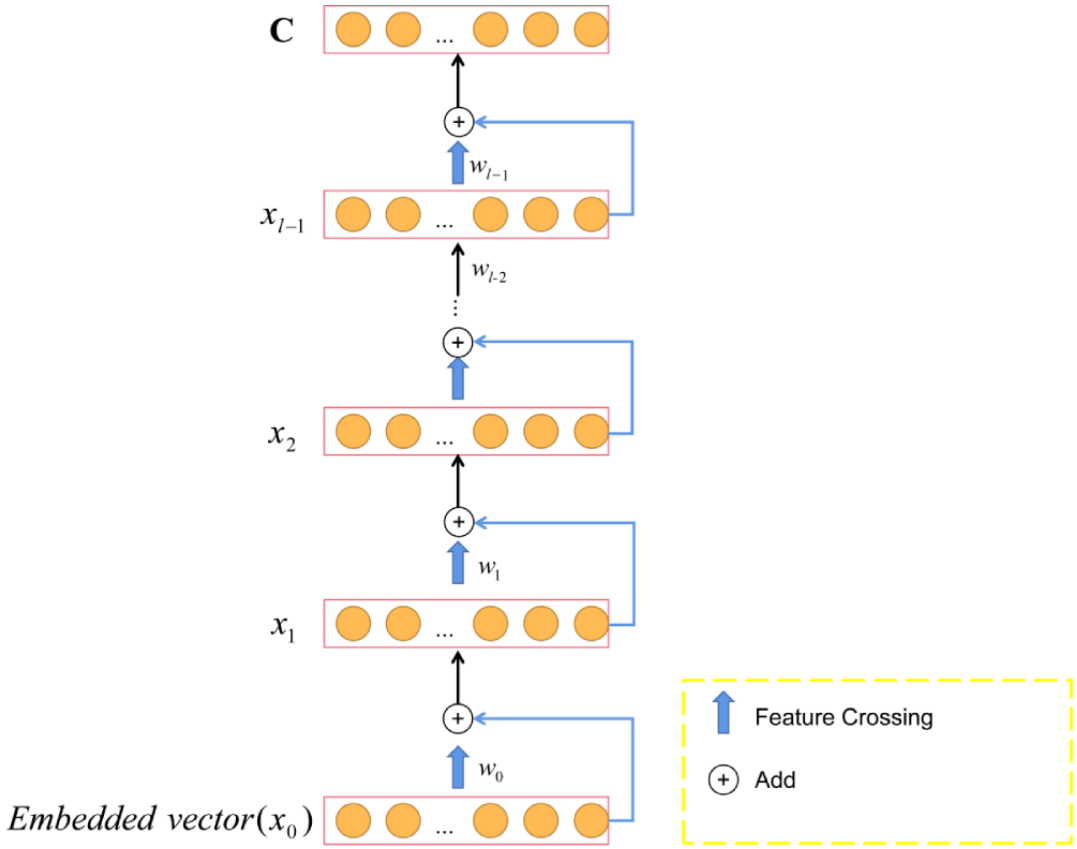
High-Order Linear Cross Feature Extraction Based on Cross Network

As shown in Figure 5, we used cross network to extract linear cross features in the model. As the number of layers increases, more cross features can be obtained. The output x_{l+1} or the layer l was obtained (1) from the original input data x_0 and the output x_l from the previous layer. $f(x_l,w_l,b_l)$ was represented by (2), where b_l represents bias, w_l represents weight, x_0 represents the initial input vector, and x_l represents the output vector of the upper layer. The advantage of (2) is that the input and output dimensions of each layer are the same while retaining the initial characteristics in each layer operation. Finally, the cross-feature vector C was obtained. Higher-order linear mixed features were obtained through the initial input vector and the previous output vector cross. The output of the previous layer was added after feature crossing, which is similar to residual and could effectively prevent gradient dissipation. As the number of layers of cross network increased, the degree of feature crossover in different fields increased.

$$x_{l+1} = f(x_l,w_l,b_l)+x_l \quad (1)$$

(2)

Figure 5. Cross network.



Multiorder Nonlinear Cross Feature Extraction Based on FGCNN Module

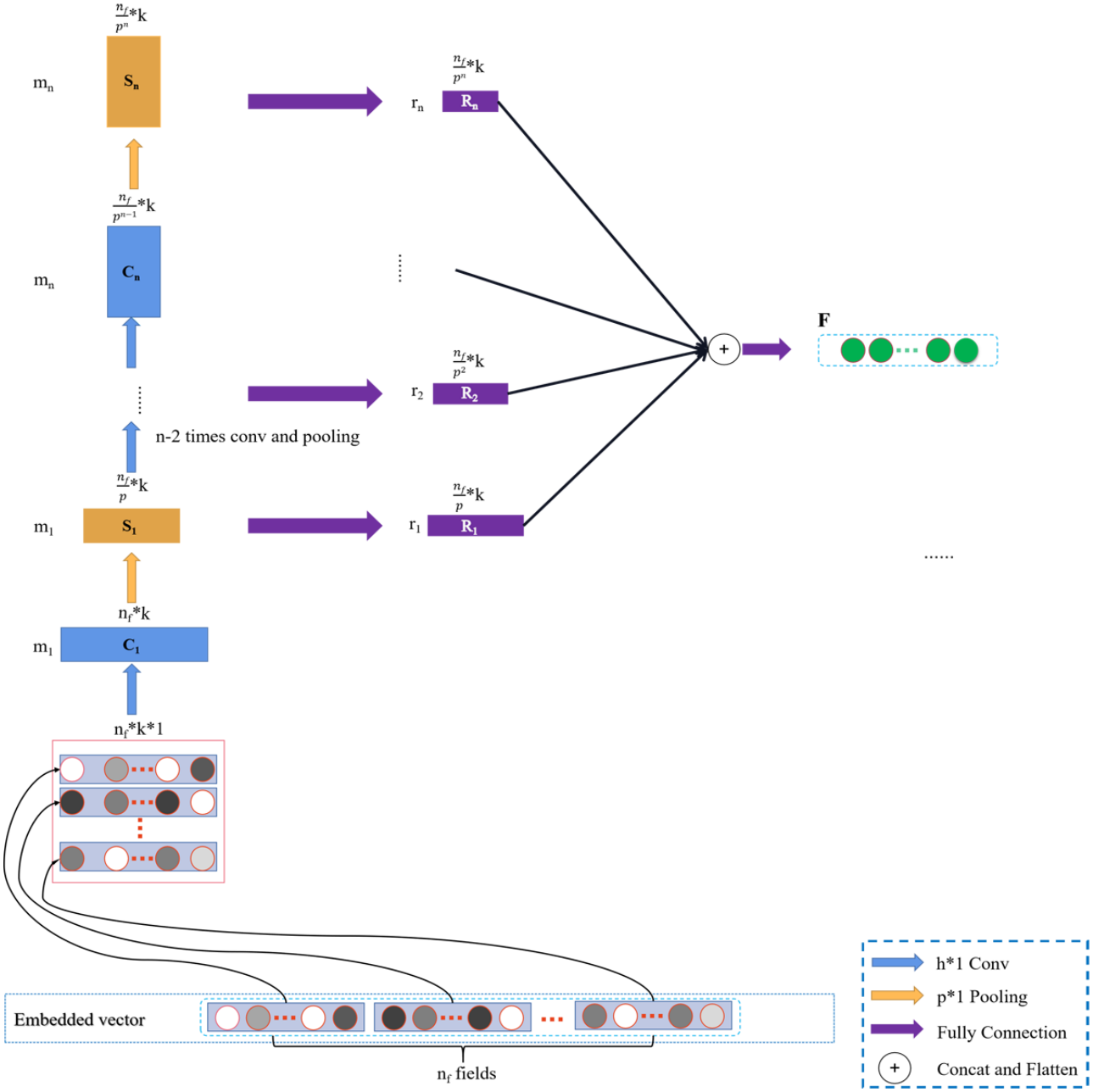
Due to the small number of parameters in the cross network, the ability of the model was limited. To capture high-order nonlinear crossover features, we introduced a deep network in parallel, as shown in Figure 6. It is difficult for the traditional neural network to learn local patterns; however, a convolution neural network can quickly obtain local patterns through convolution operation and combine it to generate a new model. Unlike text or image classification models, intelligent dialectical models have a local correlation in the input data. Therefore, after the convolution neural network output the cross features of the local patterns, it was input into an MLP model to obtain some global cross information. In the convolution layer, we entered the matrix E obtained by the embedded layer, which is an n_f*k*1 matrix, where n_f is the number of field and k is the embedding size. Then, it was convoluted with a convolution kernel with the size of h*1*m_1, and the corresponding convolution value was obtained by using tanh as the activation function. Because of the property of convolution, the convolution kernel using h*1 can obtain the cross features of adjacent h rows and output the feature graph of the specified number of channels. In terms of the first convolution, the number of channels in the convolution kernel was m_1, the size of the output convolution matrix of the first convolution C_1 was n_f*k*m_1. After obtaining the convolution matrix, the first pooled matrix S_1 was obtained through the maximum pooling layer p*1. After it was pooled, the matrix size was as follows:

The pooled matrix S_1 was passed down as the input of the next convolution layer and then recombined. Recombination was a fully connected operation shown as (3) and (4), using tanh as the activation function, B_i as the i_{th} recombination bias matrix, W_i as the i_{th} recombination weight matrix, and the first resulting high-order nonlinear feature size was as follows:

After this, repeat n-1 times convolution, pooling, and recombination, the entire convolution operation generated n reconstruction matrices, R= {R_1, R_2, …., R_n}. To better integrate with the features of cross network output, R was converted into a new matrix according to the second dimension concat according to the traditional concept of convolution network, and then the reconstruction matrix flatten was passed into an MLP. Finally, a nonlinear high-dimensional feature vector F with local and global cross features was obtained.

$$R_i = \tanh(S_i W_i + B_i) \quad (3)$$

Figure 6. Improved feature generation by convolution neural network.



Classification Module

We combined the linear and nonlinear vectors constructed automatically by FGCNN and cross network to get the vector I_1=(C,F), which is used as the input vector of the MLP. The input of the i_{th} hidden layer was O_i, and its calculation is shown in (5), where W^i represents the weight matrix of the i_{th} hidden layer and B^i represents the bias matrix of the i_{th} hidden layer.

$$O_i = \text{relu}(I_i W^i + B^i) \quad (5)$$

Since there was a multiclassification problem, we made a final decision after the last hidden layer (n_h):

Finally, in the whole cross-FGCNN model, we chose cross-entropy as the loss function of the whole model.

Where N was the total number of input data sets, Y was the real value, $\hat{y}$ was the predicted value.

Experiment

Experimental Operation

We used the cross-FGCNN and dysmenorrhea data mentioned earlier for the experiment. For training data, 75% of data were randomly selected and for test data, 25% of data were selected. A 60-dimensional label encoder vector was input, and 6-layer cross network was selected. In FGCNN, a convolution kernel with a depth of 14, 16, and 18 and a width of 4 was selected for a 3-layer convolution operation, and the depth of the MLP was

3 layers. A 3-layer neural network was selected in the classification module, and the number of neurons was 1024, 512, and 128. To maintain the robustness and optimization efficiency of the algorithm, we chose a dropout ratio of 0.2, set the learning rate to 0.001, and performed 1000 iterations.

Comparison With Other Models

A total of 6 traditional intelligent dialectical models were chosen: Bayesian classifier, multilabel K nearest neighbor (ML-KNN) classifier, 10-layer artificial neural network (ANN), decision tree, spectral clustering, and support vector machine. At the same time, to show the superiority of our algorithm for sparse symptom classification, 3 traditional CTR models DNN, FGCNN, and DCN were used in intelligent syndrome differentiation.

Results

The Experimental Results of Cross-FGCNN

Firstly, we calculated the model's accuracy, which can directly express the reliability of the model.

P represents the correct amount predicted by the model, and total represents the total number of input data from the model. The accuracy of cross-FGCNN was 96.21%. In Table 2, the amount of data between the classes of the entire data set was unbalanced, so we introduced F1 score and the confusion matrix [29].

P and R stand for precision and recall, respectively. The F1 score of cross-FGCNN was 0.9621, indicating that cross-FGCNN could be good at intelligent dialectics of TCM. Figure 7 shows the scatter diagram of the model accuracy changing with iteration times. After about 200 iterations, the accuracy of the model remained around 96%. Figure 8 shows cross-FGCNN's classification confusion matrix, where the model divided all classes into the correct classes as much as possible. In short, cross-FGCNN showed great strength in intelligent dialectical tasks.

Figure 7. Cross-FGCNN accuracy-iteration times scatter diagram. FGCNN: feature generation by convolution neural network.

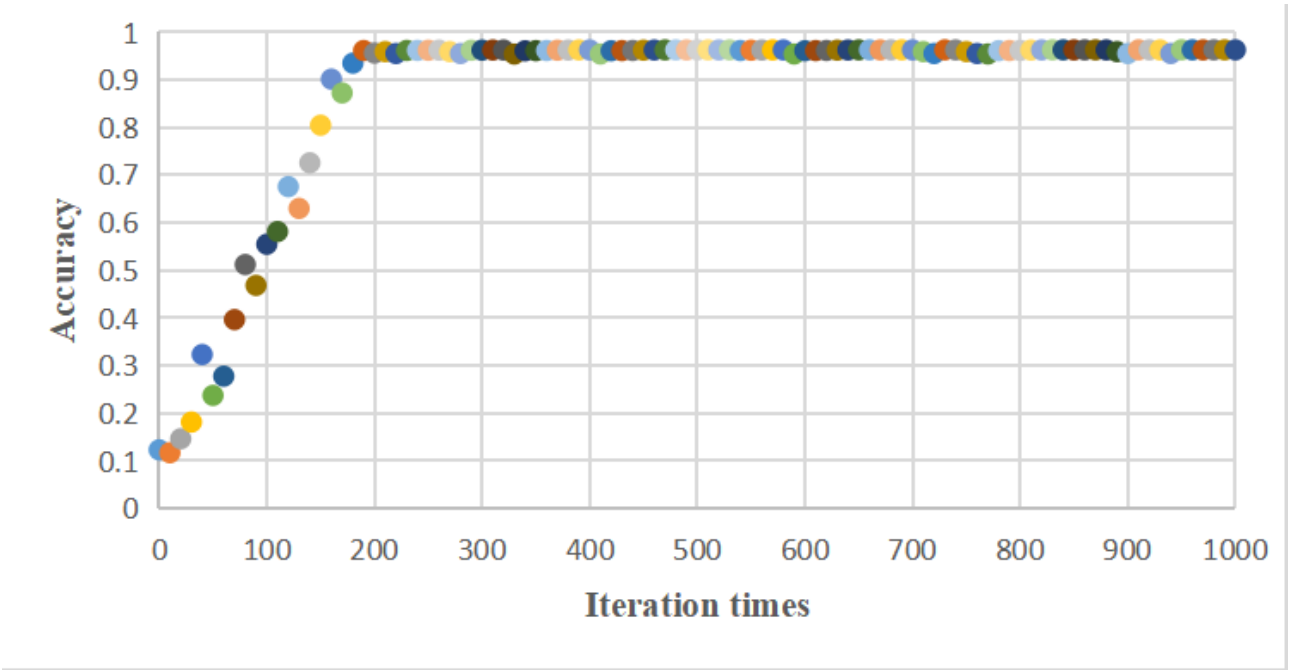
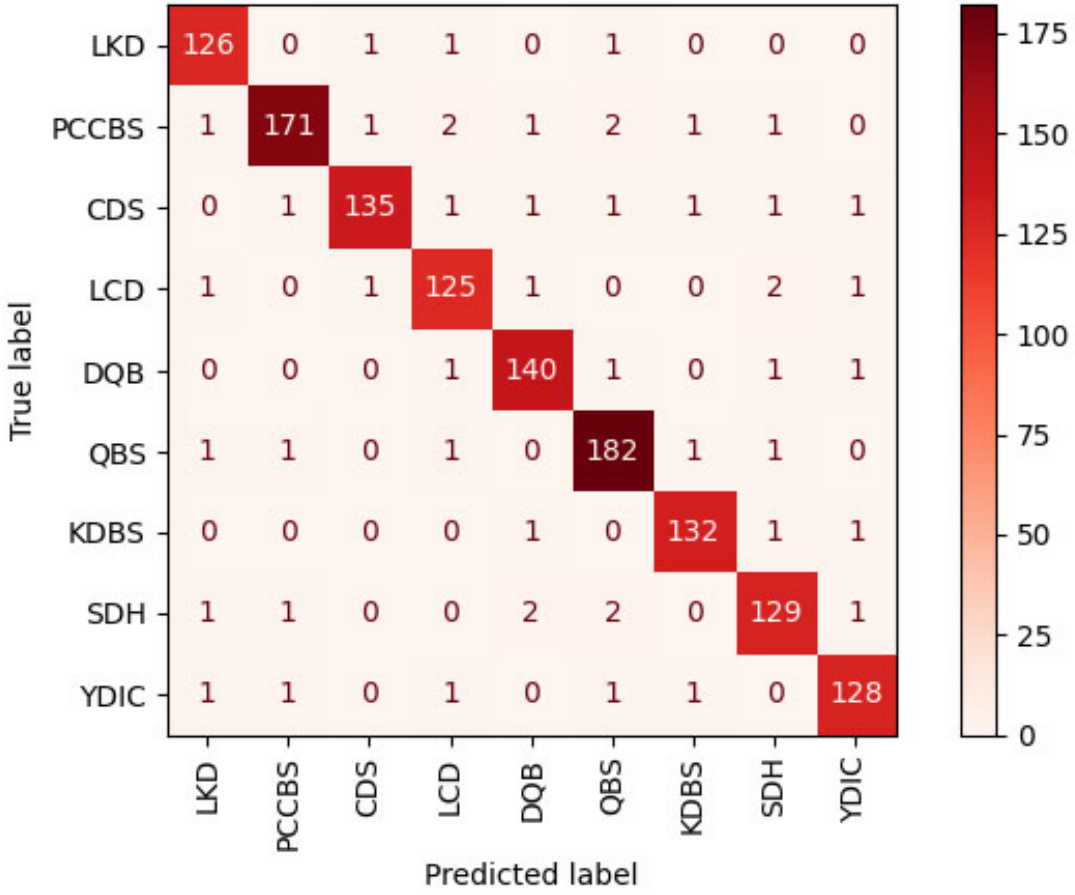


Figure 8. Cross-FGCNN confusion matrix. FGCNN: feature generation by convolution neural network; LKD: liver-kidney depletion; PCCBS: pattern of congealing cold with stasis; CDS: cold and dampness stagnation; LCD: liver constraint and dampness-heat; DQB: deficiency qi and blood; QBS: qi stagnation and blood stasis; KDBS: kidney deficiency and blood stasis; SDH: stagnation dampness-heat; YDIC: yang deficiency and internal cold.



Comparison Between Models

Table 3 shows the comparison between the cross-FGCNN model and other models with respect to accuracy and F1 score. Although 10-layer ANN [15] had a good classification effect, it still differed from the cross-FGCNN by 5% in accuracy. At the same time, the CTR model also displayed some potential in the intelligent dialectical task, where the accuracy of the FGCNN can even surpass the traditional intelligent, but there was still a gap compared with our model. To show how each model fits unbalanced classes, we introduced log-loss and receiver operating characteristic (ROC) curves.

The equation represents the calculation of log-loss: n corresponds to the number of our samples or the number of inputted instances, i corresponds to a certain sample or instance, m represents the possible number of categories of our samples, j represents a certain category, and y_{ij} denotes that for a sample i, it belongs to the label of classification j. Therefore, the smaller the log-loss, the better the fitting effect of the reaction model, and cross-FGCNN still performed well in this respect. Compared with other models, cross-FGCNN manifested great potential for intelligent dialectics, which is a high-dimensional sparse vector multiclassification task.

The abscissa of the ROC curve was a false positive rate, and the ordinate was a true positive rate. The ROC curve remained constant as the distribution of positive and negative samples in the test set changed. For TCM syndrome differentiation, some syndrome types were rare. Therefore, it is necessary to evaluate the intelligent dialectical model by ROC. Intuitively, the closer the ROC curve was to the upper left corner, the better the model classification effect was. Figure 9 and Figure 10 show the ROC curve of a new model and the traditional CTR model and the traditional intelligent dialectical model, respectively. Figure 9A-Figure 9F respectively depicts the ROC curve of decision tree, 10-layer ANN, ML-KNN, hypergraph clustering, Bayesian, and SVM, and Figure 10A-Figure 10D respectively displays the ROC curve of cross-FGCNN, deep & cross network, FGCNN, and DNN. It is clear that cross-FGCNN outperformed the other models in the classification of different syndrome types. Secondly, the area under the ROC curve can also be used as one of the indicators of the model classification effect. By comparing the area under the macroaverage ROC curve, the new intelligent dialectical model still showed great strength. For the classification comparison of single syndrome type, it is obvious that cross-FGCNN outperformed the other models in syndrome differentiation.

Table 3. Result indicators of each model.

Model	Accuracy	F1 score	Log-loss
Cross-FGCNN[a]	0.9621	0.9621	0.8356
Decision tree	0.7448	0.7439	6.4533
10-layer ANN[b]	0.9121	0.9115	1.9071
ML-KNN[c]	0.9075	0.9076	2.7211
Hypergraph clustering	0.8816	0.8814	3.8436
Bayesian	0.7816	0.7815	4.5555
SVM[d]	0.8992	0.8989	3.2289
Deep & cross network	0.7992	0.7997	3.1602
FGCNN	0.9390	0.9390	1.2820
DNN[e]	0.7220	0.6804	3.9439

[a]FGCNN: feature generation by convolution neural network.

[b]ANN: artificial neural network.

[c]ML-KNN: multilabel K nearest neighbor.

[d]SVM: support vector machine.

[e]DNN: deep neural network.

Figure 9. ROC Curves of CTR models. ROC: receiver operating characteristic; CTR: click-through-rate; FGCNN: feature generation by convolution neural network; deep neural network; LKD: liver-kidney depletion; PCCBS: pattern of congealing cold with stasis; CDS: cold and dampness stagnation; LCD: liver constraint and dampness-heat; DQB: deficiency qi and blood; QBS: qi stagnation and blood stasis; KDBS: kidney deficiency and blood stasis; SDH: stagnation dampness-heat; YDIC: yang deficiency and internal cold.

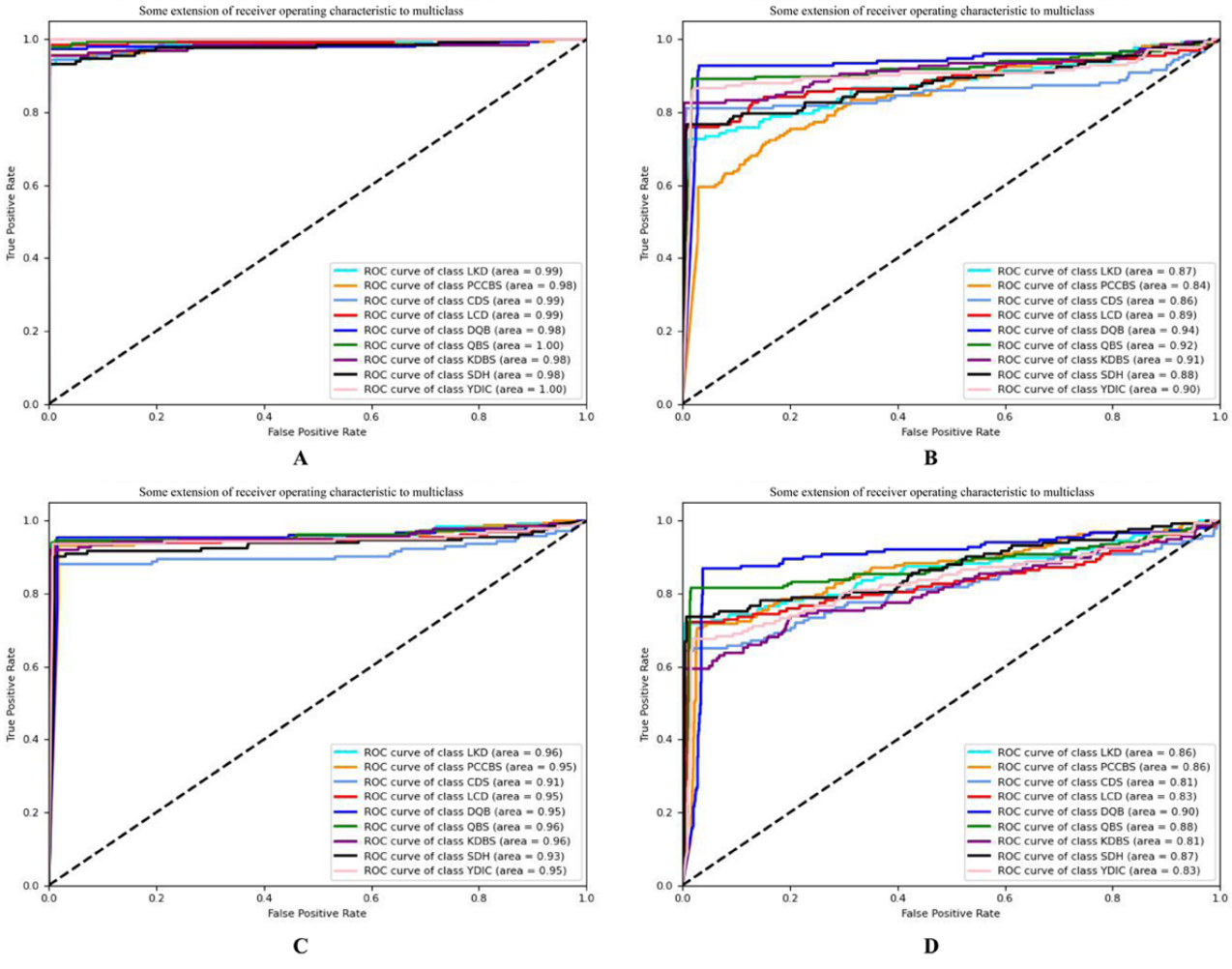
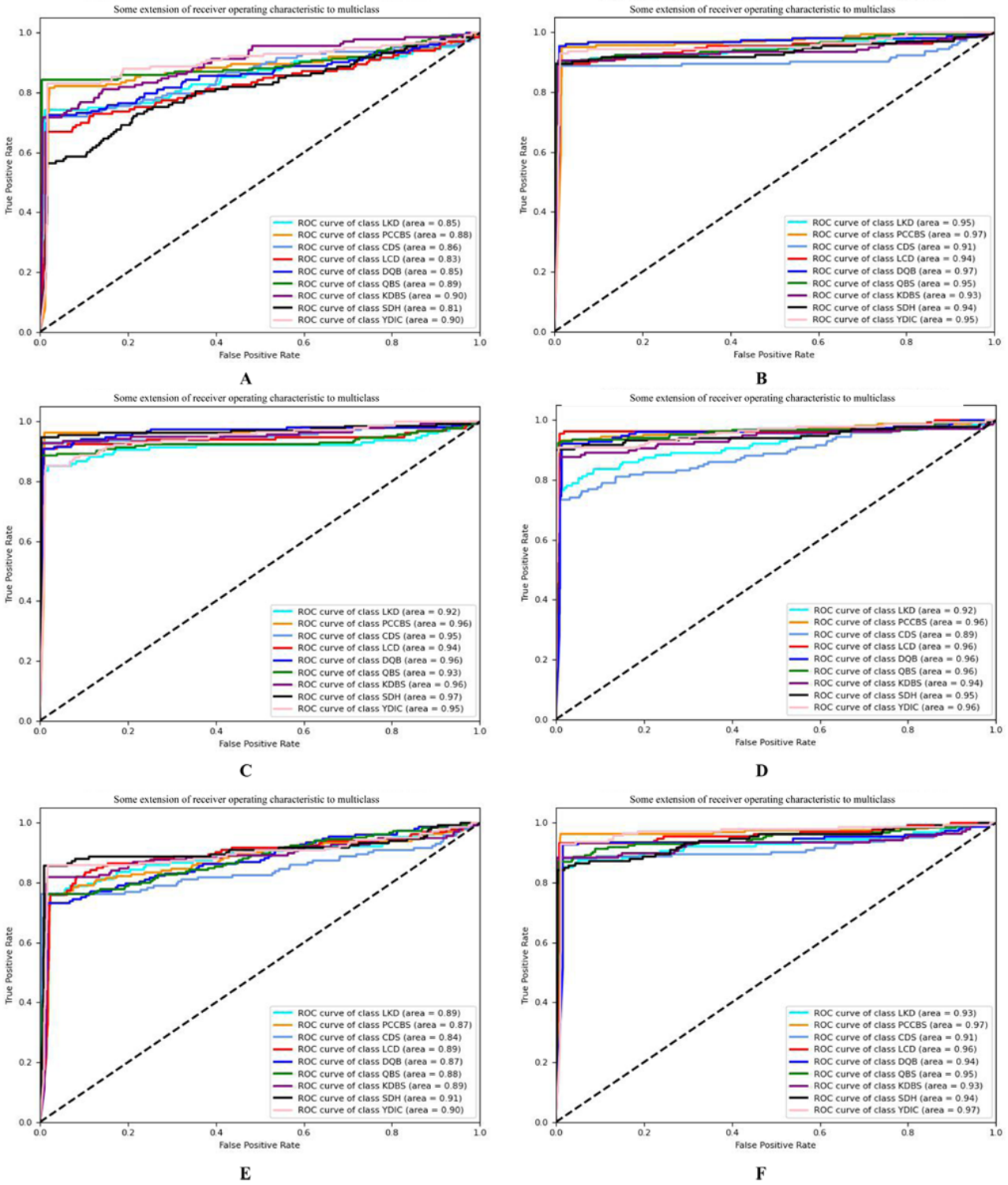


Figure 10. ROC curves of traditional intelligence dialectical models. ROC: receiver operating characteristic; ANN: artificial neural network; SVM: support vector machine; LKD: liver-kidney depletion; PCCBS: pattern of congealing cold with stasis; CDS: cold and dampness stagnation; LCD: liver constraint and dampness-heat; DQB: deficiency qi and blood; QBS: qi stagnation and blood stasis; KDBS: kidney deficiency and blood stasis; SDH: stagnation dampness-heat; YDIC: yang deficiency and internal cold; ML-KNN: multilabel K nearest neighbor.



Discussion

TCM intelligent dialectic can be regarded as a classification model of a high-dimensional sparse vector. Based on this, we improved the new intelligent dialectical model with the CTR model. Different from other related studies, this study classified different symptoms into 60 fields under the guidance of TCM diagnostics. According to this method, the expected symptom information of the 4 diagnostic methods can correspond to different fields, achieve dimensionality reduction, and standardize symptom information. This method displayed strong portability because all fields were defined following the standard of TCM, and another data set could be made to construct a new system of syndrome differentiation and treatment. Furthermore, as proof, 5273 cases of dysmenorrhea were used to train

cross-FGCNN in this study. In the model, cross network was used to construct new linear crossover vectors automatically, and FGCNN was used to construct new nonlinear crossover features. Two new features were combined for classification and compared with the other 9 types of models. Cross-FGCNN showed great potential in intelligent dialectics, a high-dimensional sparse vector multiclassification task. Nonetheless, advancements are still needed to achieve the overall optimization of the model and intelligent data acquisition.

1) The size of the data set has a great impact on the accuracy of the model, so data on dysmenorrhea and other diseases are still continuously collected to verify and improve the model.

2) The quality of data is also an essential factor affecting the model. Next, we not only continue to include new data but should also strictly check acquisition of new data and invite more professional practitioners of TCM to clean the data.

3) Intelligent medical treatment is the whole process of intelligence from data collection to patient prescription. Although we performed intelligent dialectics, it is still difficult to guarantee the reliability of data collection. There are still subjective judgments of TCM doctors in tongue diagnosis and other diagnoses, so we have begun to build an intelligent inspection model.

4) In the future, our team will construct an objective, TCM intelligent 4-diagnosis system, which integrates objective observation of TCM, intelligent listening of cough sound, remote intelligent consultation, intelligent acupoint detection of flexible portable equipment, and intelligent dialectics.

Acknowledgments

We would like to thank the Sichuan TCM big data management platform, which is the main data source. The authors obtained permission from copyright holders for reproducing or adapting any illustrations, tables, figures, or lengthy quotations previously published elsewhere. This work was supported by China National Nature Fund (81804222), National Key Research and Development Program of the Ministry of Science and Technology of China (2018YFC1707606), National Key Research and Development Program of the Ministry of Science and Technology of China (2018SZ0065), and The Popularization and Application Project of Sichuan Provincial Health Commission (20PJ168).

Conflicts of Interest

None declared.

References

1. Cyranoski D. Why Chinese medicine is heading for clinics around the world. Nature 2018 Sep;561(7724):448-450. [doi: 10.1038/d41586-018-06782-7] [Medline: 30258149]
2. Cheung F. TCM: Made in China. Nature 2011 Dec 21;480(7378):S82-S83. [doi: 10.1038/480S82a] [Medline: 22190085]
3. Chan KH, Tsoi YYS, McCall M. The effectiveness of traditional Chinese medicine (TCM) as an adjunct treatment on stable COPD patients: a systematic review and Meta-Analysis. Evid Based Complement Alternat Med 2021;2021:5550332 [FREE Full text] [doi: 10.1155/2021/5550332] [Medline: 34188688]
4. Tian J, Jin D, Bao Q, Ding Q, Zhang H, Gao Z, et al. Evidence and potential mechanisms of traditional Chinese medicine for the treatment of type 2 diabetes: a systematic review and meta-analysis. Diabetes Obes Metab 2019 Aug;21(8):1801-1816. [doi: 10.1111/dom.13760] [Medline: 31050124]
5. Liao YH, Lin JG, Lin CC, Tsai CC, Lai HL, Li TC. Traditional Chinese medicine treatment associated with female infertility in Taiwan: a population-based case-control study. Evid Based Complement Alternat Med 2020;2020:3951741 [FREE Full text] [doi: 10.1155/2020/3951741] [Medline: 33381200]
6. Lin J, Liao W, Mo Q, Yang P, Chen X, Wang X, et al. A systematic review of the efficacy comparison of acupuncture and traditional Chinese medicine in the treatment of primary dysmenorrhea. Ann Palliat Med 2020 Sep;9(5):3288-3292 [FREE Full text] [doi: 10.21037/apm-20-1734] [Medline: 33065784]
7. Xu Q, Guo Q, Wang CX, Zhang S, Wen CB, Sun T, et al. Network differentiation: a computational method of pathogenesis diagnosis in traditional Chinese medicine based on systems science. Artif Intell Med 2021 Aug;118:102134 [FREE Full text] [doi: 10.1016/j.artmed.2021.102134] [Medline: 34412850]
8. Zhang R, Zhu X, Bai H, Ning K. Network pharmacology databases for traditional Chinese medicine: review and assessment. Front Pharmacol 2019;10:123 [FREE Full text] [doi: 10.3389/fphar.2019.00123] [Medline: 30846939]
9. Liang X, Wang Q, Jiang Z, Li Z, Zhang M, Yang P, et al. Clinical research linking traditional Chinese medicine constitution types with diseases: a literature review of 1639 observational studies. J Tradit Chin Med 2020 Aug;40(4):690-702 [FREE Full text] [doi: 10.19852/j.cnki.jtcm.2020.04.019] [Medline: 32744037]
10. Xia SJ, Gao BZ, Wang SH, Guttery DS, Li CD, Zhang YD. Modeling of diagnosis for metabolic syndrome by integrating symptoms into physiochemical indexes. Biomed Pharmacother 2021 May;137:111367 [FREE Full text] [doi: 10.1016/j.biopha.2021.111367] [Medline: 33588265]

11. Zhao T, Yang X, Wan R, Yan L, Yang R, Guan Y, et al. Study of TCM syndrome identification modes for patients with type 2 diabetes mellitus based on data mining. Evid Based Complement Alternat Med 2021;2021:5528550 [FREE Full text] [doi: 10.1155/2021/5528550] [Medline: 34531918]
12. Dai L, Zhang J, Li C, Zhou C, Li S. Multi‐label feature selection with application to TCM state identification. Concurrency Computat Pract Exper 2018 Jul 30;31(23):1-13 [FREE Full text] [doi: 10.1002/cpe.4634]
13. Jiang Q, Yang X, Sun X. An aided diagnosis model of sub-health based on rough set and fuzzy mathematics: A case of TCM. IFS 2017 May 23;32(6):4135-4143 [FREE Full text] [doi: 10.3233/jifs-15958]
14. Zhang Z, Li J, Zheng W, Tian S, Wu Y, Yu Q, et al. Research on diagnosis prediction of traditional Chinese medicine diseases based on improved Bayesian combination model. Evid Based Complement Alternat Med 2021 Jun 10;2021:5513748-5513749 [FREE Full text] [doi: 10.1155/2021/5513748] [Medline: 34211562]
15. Zhou H, Li L, Zhao H, Wang Y, Du J, Zhang P, et al. A large-scale, multi-center urine biomarkers identification of coronary heart disease in TCM syndrome differentiation. J Proteome Res 2019 May 03;18(5):1994-2003. [doi: 10.1021/acs.jproteome.8b00799] [Medline: 30907085]
16. Huang WT, Hung HH, Kao YW, Ou SC, Lin YC, Cheng WZ, et al. Application of neural network and cluster analyses to differentiate TCM patterns in patients with breast cancer. Front Pharmacol 2020;11:670 [FREE Full text] [doi: 10.3389/fphar.2020.00670] [Medline: 32457636]
17. 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 [FREE Full text] [doi: 10.1109/access.2019.2921318]
18. Lin F, Xiahou J, Xu Z. TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model. Multimed Tools Appl 2016 Apr 13;75(22):14203-14232 [FREE Full text] [doi: 10.1007/s11042-016-3363-9]
19. Wang YQ, Yan HX, Guo R, Li FF, Xia CM, Yan JJ, et al. Study on intelligent syndrome differentiation in traditional Chinese medicine based on multiple information fusion methods. Int J Data Min Bioinform 2011;5(4):369-382. [doi: 10.1504/ijdmb.2011.041554] [Medline: 21954670]
20. 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 Jun 16;8(6):e17821 [FREE Full text] [doi: 10.2196/17821] [Medline: 32543445]
21. S R. Factorization Machines. 2010 Dec 1 Presented at: The 10th IEEE International Conference on Data Mining; December 14-17, 2010; Sydney, Australia. [doi: 10.1109/ICDM.2010.127]
22. S R, Z G, C F, L ST. Fast context-aware recommendations with factorization machines. 2011 Jul 24 Presented at: SIGIR '11: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieva; 2011; Beijing, China p. 635-644 URL: https://dl.acm.org/doi/pdf/10.1145/2009916.2010002 [doi: 10.1145/2009916.2010002]
23. Juan Y, Zhuang Y, Chin WS, Lin CJ. Field-aware Factorization Machines for CTR Prediction. 2016 Sep 7 Presented at: RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems; September 15-19, 2016; Boston, United States p. 43-50 URL: https://doi.org/10.1145/2959100.2959134 [doi: 10.1145/2959100.2959134]
24. Guo H, Tang R, Ye Y, Li Z, He X. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. 2017 Aug 19 Presented at: IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence; 2017; Melbourne, Australia p. 1725-1731 URL: https://doi.org//10.5555/3172077.3172127 [doi: 10.24963/ijcai.2017/239]
25. Wang R, Fu B, Fu G, Wang M. 2017 Presented at: ADKDD'17: Proceedings of the ADKDD'17; August 14, 2017; Nova Scotia, Canada URL: https://dl.acm.org/doi/10.1145/3124749.3124754 [doi: 10.1145/3124749.3124754]
26. Liu B, Tang R, Chen Y, Yu J, Zhang Y. Feature Generation by Convolutional Neural Network for Click-Through Rate Prediction. 2019 May 13 Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco p. 1119-1129 URL: https://doi.org/10.1145/3308558.3313497 [doi: 10.1145/3308558.3313497]
27. Fernández-Martínez E, Abreu-Sánchez A, Pérez-Corrales J, Ruiz-Castillo J, Velarde-García JF, Palacios-Ceña D. Living with pain and looking for a safe environment: a qualitative study among nursing students with dysmenorrhea. Int J Environ Res Public Health 2020 Sep 13;17(18):6670 [FREE Full text] [doi: 10.3390/ijerph17186670] [Medline: 32933209]
28. Kapadi R, Elander J. Pain coping, pain acceptance and analgesic use as predictors of health-related quality of life among women with primary dysmenorrhea. Eur J Obstet Gynecol Reprod Biol 2020 Mar;246:40-44. [doi: 10.1016/j.ejogrb.2019.12.032] [Medline: 31931396]
29. Ju H, Jones M, Mishra G. The prevalence and risk factors of dysmenorrhea. Epidemiol Rev 2014;36:104-113. [doi: 10.1093/epirev/mxt009] [Medline: 24284871]
30. Karout S, Soubra L, Rahme D, Karout L, Khojah HMJ, Itani R. Prevalence, risk factors, and management practices of primary dysmenorrhea among young females. BMC Womens Health 2021 Nov 08;21(1):392 [FREE Full text] [doi: 10.1186/s12905-021-01532-w] [Medline: 34749716]
31. Huang X, Su S, Duan JA, Sha X, Zhu KY, Guo J, et al. Effects and mechanisms of Shaofu-Zhuyu decoction and its major bioactive component for Cold - Stagnation and Blood - Stasis primary dysmenorrhea rats. J Ethnopharmacol 2016 Jun 20;186:234-243. [doi: 10.1016/j.jep.2016.03.067] [Medline: 27060631]

32. Armour M, Dahlen H, Smith C. More than needles: the importance of explanations and self-care advice in treating primary dysmenorrhea with acupuncture. Evid Based Complement Alternat Med 2016;2016:3467067 [FREE Full text] [doi: 10.1155/2016/3467067] [Medline: 27242909]

Abbreviations

ANN: artificial neural network
CTR: click-through-rate
DCN: deep & cross network
DNN: deep neural network
FGCNN: feature generation by convolution neural network
FM: factorization machine
ML-KNN: multilabel K nearest neighbors
MLP: multilayer perceptron
SVM: support vector machine
TCM: traditional Chinese medicine
ROC: receiver operating characteristic

Edited by C Lovis; submitted 01.04.21; peer-reviewed by W Wang, T Liyuan; comments to author 25.11.21; revised version received 17.12.21; accepted 13.02.22; published 06.04.22.

Please cite as:
Huang Z, Miao J, Chen J, Zhong Y, Yang S, Ma Y, Wen C
A Traditional Chinese Medicine Syndrome Classification Model Based on Cross-Feature Generation by Convolution Neural Network: Model Development and Validation
JMIR Med Inform 2022;10(4):e29290
URL: https://medinform.jmir.org/2022/4/e29290
doi:10.2196/29290
PMID:35384854

Original Paper

Risk Prediction of Major Adverse Cardiovascular Events Occurrence Within 6 Months After Coronary Revascularization: Machine Learning Study

Jinwan Wang[1*], MSc; Shuai Wang[2*], MM; Mark Xuefang Zhu[1], PhD; Tao Yang[3], PhD; Qingfeng Yin[4], MSc; Ya Hou[4], MSc

[1]School of Information Management, Nanjing University, Nanjing, China

[2]First Department of Cardiology, The Affiliated Hospital of Liaoning University of Traditional Chinese Medicine, Shenyang, China

[3]School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, China

[4]Jiangsu Famous Medical Technology Co Ltd, Nanjing, China

[*]these authors contributed equally

Corresponding Author:
Mark Xuefang Zhu, PhD
School of Information Management
Nanjing University
No163 Xianlin Road
Qixia District
Nanjing, 210023
China
Phone: 86 13770727298
Email: xfzhu@nju.edu.cn

Abstract

Background: As a major health hazard, the incidence of coronary heart disease has been increasing year by year. Although coronary revascularization, mainly percutaneous coronary intervention, has played an important role in the treatment of coronary heart disease, major adverse cardiovascular events (MACE) such as recurrent or persistent angina pectoris after coronary revascularization remain a very difficult problem in clinical practice.

Objective: Given the high probability of MACE after coronary revascularization, the aim of this study was to develop and validate a predictive model for MACE occurrence within 6 months based on machine learning algorithms.

Methods: A retrospective study was performed including 1004 patients who had undergone coronary revascularization at The People's Hospital of Liaoning Province and Affiliated Hospital of Liaoning University of Traditional Chinese Medicine from June 2019 to December 2020. According to the characteristics of available data, an oversampling strategy was adopted for initial preprocessing. We then employed six machine learning algorithms, including decision tree, random forest, logistic regression, naïve Bayes, support vector machine, and extreme gradient boosting (XGBoost), to develop prediction models for MACE depending on clinical information and 6-month follow-up information. Among all samples, 70% were randomly selected for training and the remaining 30% were used for model validation. Model performance was assessed based on accuracy, precision, recall, F1-score, confusion matrix, area under the receiver operating characteristic (ROC) curve (AUC), and visualization of the ROC curve.

Results: Univariate analysis showed that 21 patient characteristic variables were statistically significant (P<.05) between the groups without and with MACE. Coupled with these significant factors, among the six machine learning algorithms, XGBoost stood out with an accuracy of 0.7788, precision of 0.8058, recall of 0.7345, F1-score of 0.7685, and AUC of 0.8599. Further exploration of the models to identify factors affecting the occurrence of MACE revealed that use of anticoagulant drugs and course of the disease consistently ranked in the top two predictive factors in three developed models.

Conclusions: The machine learning risk models constructed in this study can achieve acceptable performance of MACE prediction, with XGBoost performing the best, providing a valuable reference for pointed intervention and clinical decision-making in MACE prevention.

(JMIR Med Inform 2022;10(4):e33395) doi:10.2196/33395

KEYWORDS

major adverse cardiovascular events; risk prediction; machine learning; oversampling; data imbalance

Introduction

The treatment of coronary heart disease has experienced major advances with respect to thrombolysis [1], percutaneous coronary intervention (PCI) [2], coronary artery bypass grafting (CABG) [3], and other modalities, which have significantly reduced the disability and mortality rate of coronary heart disease with increased efficacy and safety. Despite the mature use of coronary revascularization, the possible concomitant postoperative complications, including stent restenosis, stent thrombosis, coronary microvascular dysfunction, myocardium ischemic/reperfusion injury, depression/anxiety before and after surgery, and procedure-related vascular complications, have led to a high rate of major adverse cardiovascular events (MACE), with an incidence of approximately 15% to 25%, mainly occurring within 6 months or 12-18 months after the operation [4,5]. The occurrence of MACE is a serious issue that markedly affects the prognosis of patients; thus, developing methods to reduce or even avoid MACE has been a long-standing and imperative clinical challenge. Faced with these needs, a reliable risk prediction model of MACE after coronary revascularization can effectively predict the severity of disease to help clinicians and patients in the shared decision-making process of treatment and rehabilitation plans, which is of practical significance to take early measures so that interventions can be delivered early to reduce the probability of adverse events.

In recent years, there has been an explosion of studies on MACE risk assessment, which can be divided into rule-based expert systems, statistical-based analysis techniques, and machine learning (ML)-based prediction models [6]. As a representative expert system, the assistive diagnostic system MYCIN, developed by Shortliffe et al [7], uses predicate logic and first-order logic to imitate the reasoning process of an expert to identify bacterial infections and provide available treatment options. However, this medical expert system requires a manual summarization of a large number of expert rules, which leads to high maintenance costs and poor expansibility. In response to these problems, statistical analysis has been incorporated into medical data processing to aid clinical decision-making by exploring the relationship between target and explanatory variables [8,9]. With the continuous development of data mining, ML algorithms [10] have been gradually applied in the field of clinical medical research [11,12] by virtue of the powerful data processing and knowledge representation capabilities, achieving better predictive performance by deeply mining the inherent laws of data to obtain insight into the tendency of future development. Disease identification and prediction are often regarded by ML as a classification problem with clinical manifestations as feature variables and the corresponding diagnostic results as targeted labels. For example, Zhu et al [13] constructed a model for predicting the risk of central lymph node metastasis utilizing available preoperative characteristics and intraoperative frozen section information. Patel et al [14] developed a fast and efficient detection technique for heart disease based on 303 records with 76 attributes. Duan et al [15] proposed a novel approach of MACE prediction for patients with acute coronary syndrome using not only static patient features but also dynamic treatment information during their hospitalization, which appeared to boost the performance and readily meet the clinical prediction demand.

These studies have indicated that ML has better predictive performance over conventional statistical approaches. Hence, it might be a better choice to develop a predictive model by capitalizing on the strong generalization and robustness of ML methods. However, in clinical reality, a nonnegligible problem is that the distribution of data is often imbalanced and can even be severely imbalanced in some cases [16]; that is, the number of samples with MACE occurrence is significantly smaller than that without MACE occurrence. In such a case, poor risk models may be obtained because the decision boundary is likely biased in response to the unbalanced data [17].

Data imbalance is a common clinical occurrence [18]. For example, in early cancer screening, the general population is much larger than the population of cancer patients [19]. In the identification of frailty in the elderly, the number of subjects from the negative sample far exceeds that of the positive sample [20]. Similarly, in considering risk prediction of MACE occurrence after coronary revascularization, there are relatively fewer patients with MACE occurrence than without. The challenge with using imbalanced data sets is that most ML techniques, which aim for overall classification accuracy, will ignore the minority class in model training, making the minority perform poorly [21]. In such a case, although high overall accuracy can be achieved, the recognition rate of the minority class is extremely low, which is usually more important. Sample reconstruction is a commonly used intervention for imbalanced classification to balance the positive and negative classes, mainly including undersampling and oversampling [22]. Undersampling aims to balance uneven data sets by removing data from the majority class and keeping all of the data in the minority class. Although it is a common and important approach, undersampling can somewhat affect the model performance as some potentially important information can be lost. Conversely, oversampling extends the size of the minority class by duplicating or synthesizing. This approach is appropriate when the original sample set does not contain sufficient information. The most frequently employed oversampling approach is the synthetic minority oversampling technique (SMOTE) [23], which has been successfully applied in imbalanced learning in various fields [24,25], including clinical research [26,27]. For example, Ishaq et al [27] proved that their model achieved the best performance on a data set that was balanced with the SMOTE technique in the prediction of survival for patients with heart disease. In addition to direct employment of the original SMOTE technique, some improved versions have been developed to

synthesize higher-quality samples [28,29]. For example, Prusty et al [28] proposed the weighted-SMOTE approach, in which oversampling of each minority data sample is carried out based on an assigned weight.

Accordingly, to achieve risk prediction of MACE after coronary revascularization, the aim of this study was to establish prediction models using ML algorithms based on sufficient data processing. First, the SMOTE technique was adopted to balance the initial imbalanced data set. For model construction, six algorithms were respectively employed to build six predictive models, and then the optimal model was determined according to systematic comparison and evaluation. The models were then further explored to identify factors affecting the occurrence of MACE. This study can therefore provide a valuable reference for pointed intervention and clinical decision-making in MACE prevention.

Methods

Study Participants

We retrospectively collected the medical records of patients who underwent coronary revascularization at The People's Hospital of Liaoning Province and Affiliated Hospital of Liaoning University of Traditional Chinese Medicine from June 2019 to December 2020, including clinical information and follow-up information within 6 months of surgery.

Inclusion and Exclusion Criteria

The general inclusion criteria were as follows: (1) age ≥18 years and ≤85 years; (2) patients with previous coronary revascularization (including CABG and/or PCI).

Exclusion criteria were as follows: (1) patients with incomplete medical records and unable to provide original surgical information; (2) patients who had not undergone coronary revascularization or for whom the surgery failed; (3) patients who required mechanical assistive therapy with an intraaortic balloon pump (IABP) after successful coronary revascularization treatment, since these patients are critically ill, requiring IABP treatment to maintain vital signs and do not have indications for discharge or follow-up; (4) combined with other heart diseases such as malignant arrhythmia, cardiac insufficiency before and after surgery (ie, patients with New York Heart Association class IV or Killip class IV), or severe cardiopulmonary insufficiency, as these patients are in a severe condition and have underlying diseases resulting in a poor prognosis or even surgical intervention, leading to lack of follow-up or are already at the endpoint before enrollment; and (5) patients with neuropathy or those who may not be able to participate in the study due to literacy, language, or other communication barriers.

Data Exploration

Before data modeling, data analysis and preprocessing are indispensable [30]. We used the occurrence of MACE within 6 months after coronary revascularization as the study endpoint. For this study, MACE was defined to involve all-cause deaths, nonfatal myocardial infarction, recurrent angina, repeat revascularization, stroke, and readmission within 6 months after coronary revascularization. The total number of characteristic variables in the raw data set was 49, which mainly involved five aspects: subject characteristics, medical history, drug prescriptions, clinical events, and clinical psychiatric evaluations. Initially, removing the records with null clinical endpoints and those with more than 80% missing features, we obtained a data set containing 1004 records, including 753 without MACE and 251 with MACE. Subsequently, eight unimportant characteristic variables with high missing rates (over 60%) were deleted through communication with clinical experts, and the missing values of the remaining 41 characteristic variables, if any, were filled in. The specific data-filling approach was as follows. First, we logged into the Data Management Center of Jiangsu Famous Medical Technology Co Ltd, the cloud storage platform for the data, to search for missing values, because there may have been a system failure during the data export process. If not available, we would continue to search for paper copies of the original information and records. If these were not found, data-filling methods were employed [31] using the expectation-maximization algorithms for continuous variables and the mode for filling in missing data of discrete variables.

The preliminary exploratory analysis revealed that the original data set had a category imbalance problem; that is, the ratio of the number of samples with and without MACE was approximately 1:3, which would affect the performance of the final risk prediction model to a certain extent. Therefore, we determined that the original data set should first be processed by equalization. Currently, the main strategies to solve the imbalanced classification problem include oversampling and undersampling [32], among which the former, represented by SMOTE [23,33], is widely believed to be an effective strategy for resolving class imbalance. Therefore, we adopted SMOTE for sample reconstruction in this study.

SMOTE Technique

SMOTE is a novel oversampling technique proposed by Chawla et al [23], which has become an effective preprocessing technique for uneven data sets. In contrast to many traditional oversampling methods, SMOTE does not simply duplicate the samples but rather increases the number in the minority class by creating new synthetic samples. This reduces the likelihood of overfitting and improves the generalization performance of the classifier on the test set. The algorithm flow is as follows [23]:

(1) For each sample x in the minority class, calculate the Euclidean distance between x and all samples in the minority class and obtain its k-nearest neighbors.

(2) Select several samples from the k-nearest neighbors of x at random.

(3) For each randomly selected neighbor x_n, a new sample is synthesized according to the formula:

$$x_{new} = x + rand(0,1) \times (x_n - x)$$

Depending on the sampling rate, we set the execution time and repeated the above process. Finally, we obtained the final

minority class by combining the synthetic samples with the original samples.

Research Technique

All data were statistically analyzed with SPSS 26.0 software. The enumeration data are expressed as count (percentage), processed with a χ^2 test, whereas the measurement data are presented as means (SD) and analyzed by *t* tests. A *P* value less than .05 was accepted to indicate statistical significance.

ML algorithms are characterized by better performance compared with traditional statistical methods in risk prediction, which were selected for modeling in MACE prediction in this study. We randomly separated the entire data set into a training set and validation set with an approximate ratio of 7:3, in which the training set was used to construct the prediction model and the validation set was used to verify and evaluate the model performance. Six ML algorithms were employed to construct risk prediction models: decision tree (DT), random forest (RF), logistic regression (LR), naïve Bayes (NB), support vector machine (SVM), and extreme gradient boosting (XGBoost). Among them, RF and XGBoost are ensemble ML classifiers and the others are single classifiers. Throughout the experiment, we implemented modeling and evaluation using Python 3.8 with open-source Python libraries. During the training process, the optimal parameters were determined by 10-fold cross-validation to prevent overfitting, and then we obtained the final ML-based risk models of MACE prediction.

Evaluation Metrics

The performance of ML models is often assessed with certain evaluation metrics [34]. The blend of various evaluation metrics is expected to facilitate analytical research [35]. In this study, the indicators accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve (AUC) were all employed for model evaluation. Values closer to 1 for these metrics indicate better performance of the predictive models. We also used the ROC curve as a common measure to graphically visualize the discriminative power of models.

For classification tasks, the confusion matrix [34] is also a critical index in model evaluation. The confusion matrix for binary classification is shown in Table 1.

Based on the confusion matrix, the values of the other evaluation metrics can be readily calculated, as follows:

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

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

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

$$\text{F1-score}=2\times TP/(2\times TP+FP+FN)=2\times \text{precision}\times \text{recall}/(\text{precision}+\text{recall})$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Table 1. The confusion matrix for binary classification.

Labeled	Predicted as negative	Predicted as positive
Negative	True negative	False positive
Positive	False negative	True positive

Ethics Approval

This study was approved by the Institutional Review Board of The Affiliated Hospital of Liaoning University of Traditional Chinese Medicine (2019034FS(KT)-016-02).

Results

Univariate Analysis

Based on expert experience, a total of 1004 samples with 41 characteristic variables were finally adopted for model construction after data preprocessing, including 251 cases with MACE and 753 cases without MACE. The detailed statistical information of the feature variables and results of the univariate analysis of MACE are shown in Table 2. Due to space limitations, only the statistically significant characteristic variables are presented, and the complete information of the 41 variables is provided in Multimedia Appendix 1. The results revealed that 21 characteristics were significantly different ($P<.05$) between the groups without and with MACE, namely age, smoking, work, course of the disease, family history, seasonal onset, previous myocardial infarction, dyslipidemia, brain infarction, cardiac insufficiency, traditional Chinese medicine (TCM) treatment, anticoagulant drugs, antiarrhythmic drugs, diuretic, lansoprazole injection, bleeding events, left atrial diameter (LAD), left ventricular ejection fraction (LVEF), bypass surgery, Hamilton anxiety scale (HAMA), and Hamilton depression scale (HAMD).

Table 2. Significant variables in univariate analysis.

Characteristics	Without MACE[a] (n=753)	With MACE (n=251)	Statistic[b]	*df*	*P* value
Age (years), mean (SD)	63.47 (10.98)	66.82 (11.10)	–4.180	1002	<.001
Smoking, n (%)			4.360	1	.04
No	429 (57.0)	124 (49.4)			
Yes	324 (43.0)	127 (50.6)			
Type of work, n (%)			6.213	1	.013
Physical work	353 (46.9)	95 (37.8)			
Mental work	400 (53.1)	156 (62.2)			
Course of disease (years since diagnosis), mean (SD)	3.41 (5.11)	5.43 (5.81)	–4.930	387.17	<.001
Family history, n (%)			4.387	1	.04
No	693 (92.0)	220 (87.6)			
Yes	60 (8.0)	31 (12.4)			
Seasonal onset, n (%)			17.920	1	<.001
No obvious seasonality	675 (89.6)	199 (79.3)			
Obvious seasonality	78 (10.4)	52 (20.7)			
Previous myocardial infarction, n (%)			80.775	1	<.001
No	643 (85.4)	147 (58.6)			
Yes	110 (14.6)	104 (41.4)			
Dyslipidemia, n (%)			6.659	1	.01
No	681 (90.4)	240 (95.6)			
Yes	72 (9.6)	11 (4.4)			
Brain infarction, n (%)			24.822	1	<.001
No	671 (89.1)	192 (76.5)			
Yes	82 (10.9)	59 (23.5)			
Cardiac insufficiency, n (%)			6.249	1	.01
No	712 (94.6)	226 (90.0)			
Yes	41 (5.4)	25 (10.0)			
TCM [c] treatment, n (%)			4.489	1	.03
No	570 (75.7)	173 (68.9)			
Yes	183 (24.3)	78 (31.1)			
Anticoagulant drugs, n (%)			47.408	1	<.001
No	367 (48.7)	185 (73.7)			
Yes	386 (51.3)	66 (26.3)			
Antiarrhythmic drugs, n (%)			4.123	1	.04
No	703 (93.4)	243 (96.8)			
Yes	50 (6.6)	8 (3.2)			
Diuretic, n (%)			6.055	1	.01
No	636 (84.5)	195 (77.7)			
Yes	117 (15.5)	56 (22.3)			
Lansoprazole injection, n (%)			14.381	1	<.001
No	634 (84.2)	235 (93.6)			
Yes	119 (15.8)	16 (6.4)			

Characteristics	Without MACE[a] (n=753)	With MACE (n=251)	Statistic[b]	*df*	*P* value
Bleeding events, n (%)			12.446	1	<.001
No	735 (97.6)	233 (92.8)			
Yes	18 (2.4)	18 (7.2)			
LAD[d] (mm), mean (SD)	36.59 (4.91)	37.70 (5.54)	–2.988	1002	.003
LVEF[e] (%), mean (SD)	52.65 (8.08)	51.31 (8.91)	2.113	395.97	.04
Bypass surgery, n (%)			7.200	1	.007
No	745 (98.9)	242 (96.4)			
Yes	8 (1.1)	9 (3.6)			
HAMD[f], mean (SD)	7.23 (5.26)	9.27 (5.87)	–4.877	392.12	<.001
HAMA[g], mean (SD)	8.23 (6.59)	11.13 (6.83)	–5.979	1002	<.001

[a]MACE: major adverse cardiovascular events.

[b]*t* statistics for continuous variable comparisons and χ^2 statistics for categorical variables.

[c]TCM: traditional Chinese medicine.

[d]LAD: left atrial diameter.

[e]LVEF: left ventricular ejection fraction.

[f]HAMD: Hamilton depression scale.

[g]HAMA: Hamilton anxiety scale.

Oversampling

To cope with data imbalances in the original data sets, the SMOTE algorithm was employed. The sample distribution before and after oversampling is shown in Table 3. The ratio of sample numbers with and without MACE occurrence was 1:1 after oversampling for both the training and validation set.

Table 3. Data distribution before and after oversampling.

Oversampling	Training set		Validation set	
	Without MACE[a]	With MACE	Without MACE	With MACE
Before	527	176	226	75
After	527	527	226	226

[a]MACE: major adverse cardiovascular events.

Modeling and Evaluation

Taking whether MACE occurred within 6 months as the label and 21 statistically significant factors in the univariate analysis as features, the MACE risk prediction models were constructed by DT, RF, LR, NB, SVM, and XGBoost, respectively. As the central aspect, model evaluation is quite essential. First, we comprehensively compared ML algorithms before and after oversampling to test the effectiveness of the SMOTE strategy, with specific results presented Table 4. The performance of the ML models based on the oversampled data set was significantly better than that of models based on the original imbalanced dataset, thus demonstrating the rationality of the oversampling strategy. It is worth noting that although the accuracy before oversampling was slightly higher than that obtained after oversampling, other indicators such as precision, recall, F1-score, and AUC were significantly lower than those obtained after oversampling, especially precision, recall, and F1-score. The reason for the high accuracy before oversampling is that this comes at the expense of the accuracy of minority samples to improve the overall accuracy, which is of little significance for the imbalanced classification problem [36], whereas the evaluation indicators such as precision, recall, F1-score, and AUC should be more relevant than the overall accuracy on imbalanced issues.

The ROC curves for the six ML algorithms based on balanced data sets are detailed in Figure 1. Combined with the results of Table 4, it is clear that XGBoost and RF had better performance with respect to accuracy, precision, F1-score, and AUC, with XGBoost having the best effect. This is likely because both XGBoost and RF belong to ensemble learning methods, with the advantages of integrating the performance of multiple weak classifiers. However, NB outperformed the other models in terms of recall. From the definition, as detailed above, recall is the proportion of correctly identified positive samples among all positive samples, which indicates that NB is more sensitive to positive samples than other models. However, as previously explained, no single indicator exists that can comprehensively evaluate a model's performance. NB had the lowest values for all metrics except for recall. Therefore, it is clear that XGBoost achieved optimal performance in MACE prediction from an overall perspective, with an accuracy of 0.7788, precision of

0.8058, recall of 0.7345, F1-score of 0.7685, and AUC of 0.8599.

As another method to assess the effectiveness of classification, the confusion matrices of all methods are illustrated in Figure 2. Specifically, 0 stands for the negative samples (ie, patients without MACE occurrence) and 1 represents the positive samples (ie, patients with MACE occurrence). It can be intuitively seen that DT, RF, SVM, and XGBoost had higher recognition rates for negative samples, with XGBoost performing the best. The fact that XGBoost had only 100 misclassified samples, which was the lowest among all models, further proving its superiority. In contrast, LR and NB had higher identification rates for positive samples, with NB accurately identifying 192 positive samples, confirming that NB is more sensitive to the minority class. However, NB only identified 127 negative samples, which was the lowest recognition rate of negative samples among all models. In addition, there were 133 misclassified samples with NB, which was second only to DT. Combining these results with those shown in Table 4, we can infer that NB is poor at identifying negative samples despite its high recall, which suggests that the classification boundary is biased toward the minority class (ie, patients with MACE).

For a deeper exploration and interpretation of the constructed models, the relative importance of feature variables in each MACE-predicting model is shown in descending order in Figure 3. Since the SVM prediction model used in this study adopted the radial basis function—a complex Gaussian kernel function that makes the SVM model a black box—the direct influence of each feature variable on the SVM model could not be obtained. Similarly, the algorithm of NB used in this study was Gaussian NB. Therefore, only the results of the other ML prediction models are shown.

As shown in Figure 3, the overall trends of DT, RF, and XGBoost demonstrated similar performance, although the relative importance rankings of the three ML models were not completely consistent. More specifically, anticoagulant drugs and the course of disease consistently ranked in the top 2 for all three prediction models. In contrast, the top 2 important features of LR were previous myocardial infarction and HAMA. The relative importance of high-ranking features of XGBoost, the optimal model as a whole, was as follows (in descending order): anticoagulant drugs, course of the disease, smoking, lansoprazole injection, dyslipidemia, HAMA, diuretic, LAD, seasonal onset, bleeding events, HAMD, LVEF, age, antiarrhythmic drugs, TCM treatment, previous myocardial infarction, brain infarction, work, and cardiac insufficiency.

Table 4. Comparisons of machine learning algorithms before and after oversampling.

Algorithms	Accuracy	Precision	Recall	F1-score	AUC[a]
Before oversampling					
DT[b]	0.7575	0.5217	0.3200	0.3967	0.7296
RF[c]	0.7741	0.6667	0.1867	0.2917	0.7888
LR[d]	0.7608	0.5405	0.2667	0.3571	0.7534
NB[e]	0.7442	0.4857	0.4533	0.4689	0.7224
SVM[f]	0.7641	0.7	0.0933	0.1647	0.7431
XGBoost[g]	0.7807	0.5918	0.3867	0.4677	0.7873
After oversampling					
DT	0.7035	0.73	0.6460	0.6854	0.7748
RF	0.7522	0.7714	0.7168	0.7431	0.8434
LR	0.7434	0.7254	0.7832	0.7532	0.7841
NB	0.7058	0.6598	0.8495	0.7421	0.7463
SVM	0.7478	0.7593	0.7257	0.7421	0.8075
XGBoost	0.7788	0.8058	0.7345	0.7685	0.8599

[a]AUC: area under the curve.

[b]DT: decision tree.

[c]RF: random forest.

[d]LR: logistic regression.

[e]NB: naïve Bayes.

[f]SVM: support vector machine.

[g]XGBoost: extreme gradient boosting.

Figure 1. ROC curves of machine learning algorithms after oversampling. ROC: receiver operating characteristic; DT: decision tree; RF: random forest; LR: logistic regression; NB: naïve Bayes; SVM: support vector machine; XGBoost: extreme gradient boosting; TPR: true positive rate; FPR: false positive rate.

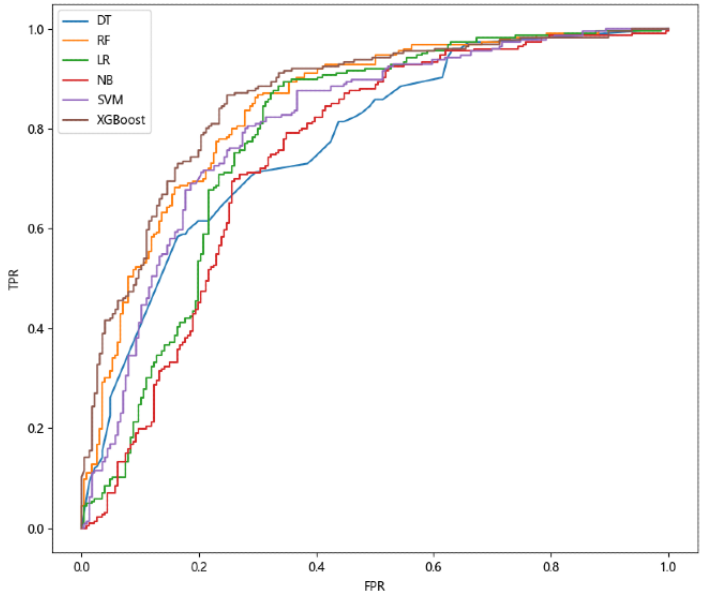


Figure 2. Confusion matrix of the risk prediction models with machine learning algorithms: (A) decision tree (DT), (B) random forest (RF), (C) logistic regression (LR), (D) naïve Bayes (NB), (E) support vector machine (SVM), (F) extreme gradient boosting (XGBoost).

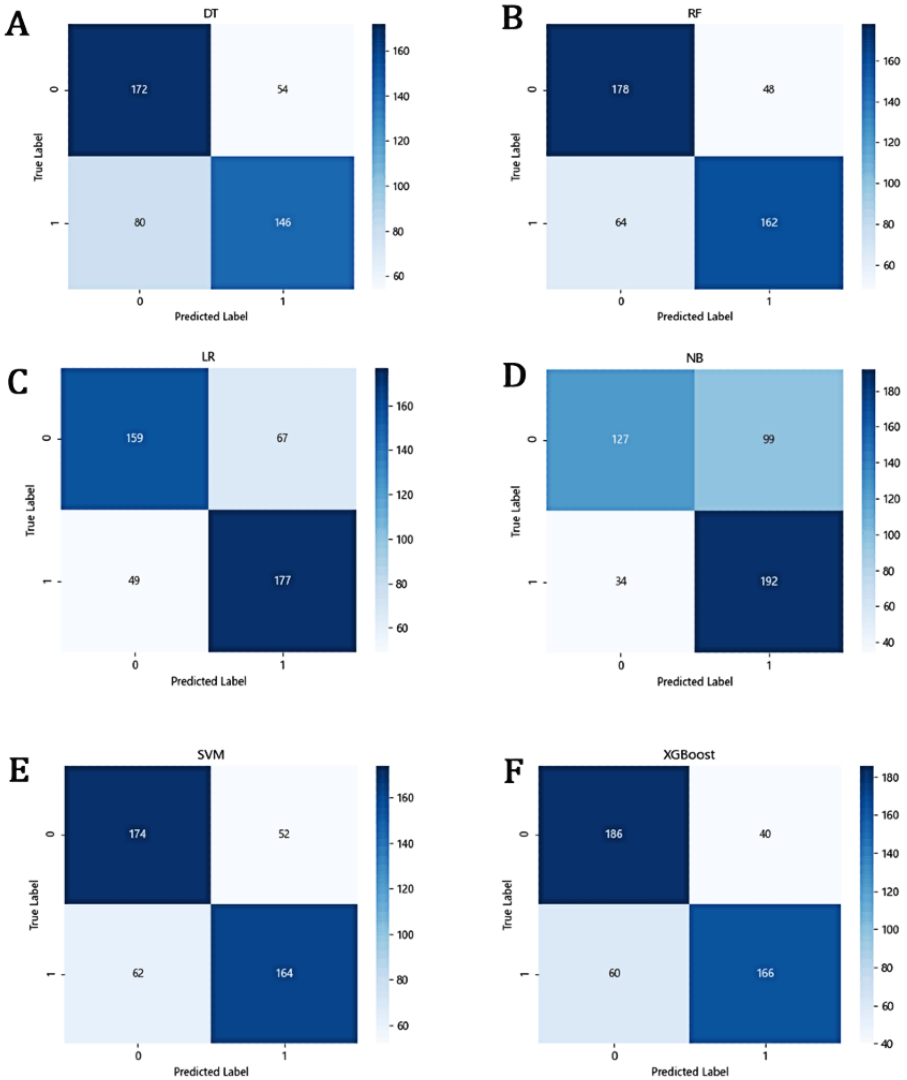
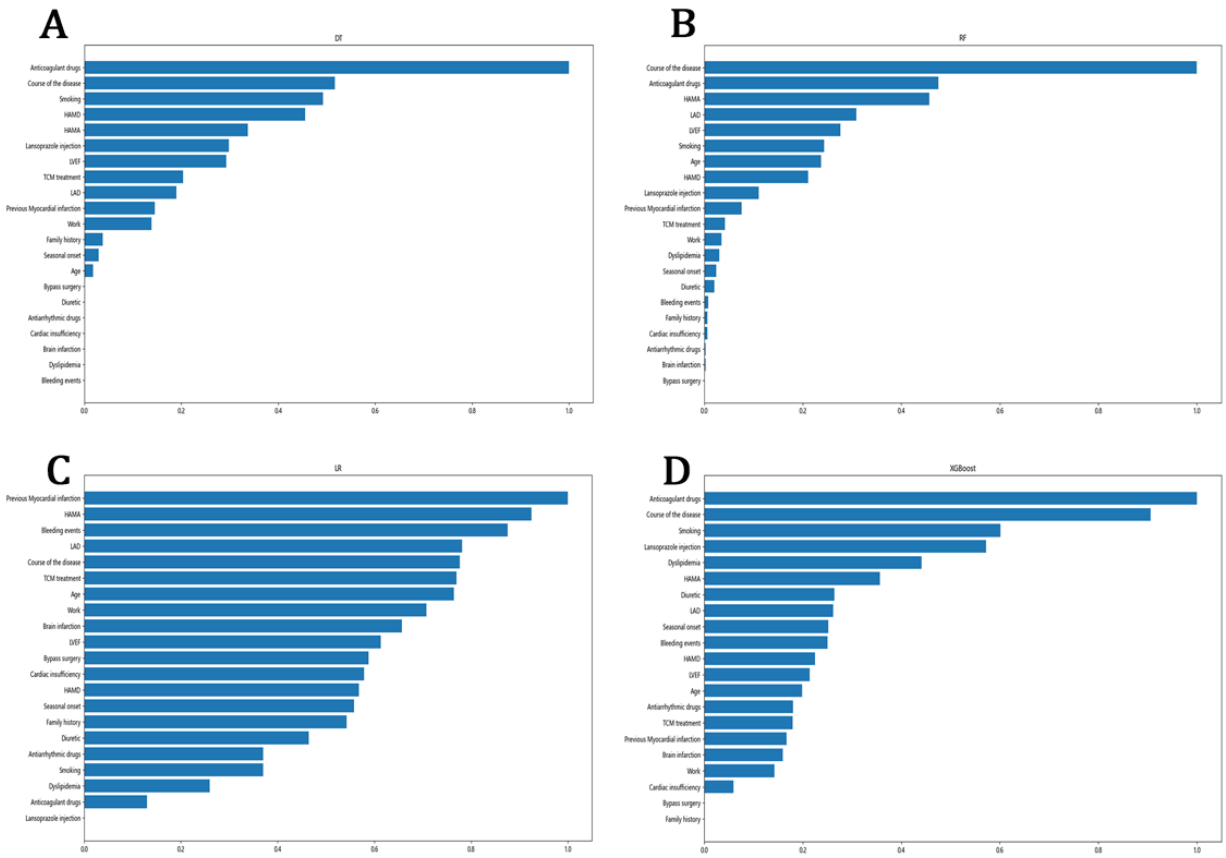


Figure 3. The relative importance of feature variables of the risk prediction models with machine learning algorithms: (A) decision tree (DT), (B) random forest (RF), (C) logistic regression (LR), (D) extreme gradient boosting (XGBoost). TCM: traditional Chinese medicine; LAD: left atrial diameter; LVEF: left ventricular ejection fraction; HAMD: Hamilton depression scale; HAMA: Hamilton anxiety scale.



Discussion

Principal Results

MACE such as recurrent angina can still occur after coronary revascularization, thus affecting the efficacy and prognosis of surgery. Detecting the postoperative characteristics of patients and combining them with preoperative information to establish a risk assessment model can provide timely warning of the risk of MACE occurrence, thereby helping medical staff and patients to intervene in a timely manner and achieve the purpose of treating the disease before it occurs [37]. In this study, we constructed and evaluated multiple risk models with ML algorithms for MACE prediction in patients within 6 months after coronary revascularization. Performance comparisons of the ML models demonstrated that the XGBoost model performed the best from an overall perspective. Moreover, a deeper exploration of the relative importance of feature variables of the constructed ML models was performed, which is valuable to provide a reference for the pointed intervention and clinical decision-making in MACE prevention.

According to existing studies, the risk factors of MACE after coronary revascularization can be roughly divided into two categories [38]: (1) uncontrollable factors such as gender, age, and family history; and (2) controllable factors such as environment and personal undesirable lifestyle habits. The finding that the risk of MACE occurrence increases with age is similar to that of previous studies [39]. The American Heart Association lists seven major controllable risk factors for coronary heart disease [40]: smoking, physical inactivity, diet, being overweight or obese, abnormal cholesterol levels, high blood pressure, and diabetes. Ritchie et al [41] demonstrated that environment and personal habits contribute to a higher risk of MACE occurrence. Likewise, we found that smokers had a 5.8% higher MACE incidence than that of nonsmokers and mental workers had a 6.9% higher rate than that of manual workers, possibly due to lack of exercise. In addition, we discovered that the occurrence of MACE was correlated with the course of the disease and seasonal changes, which is in line with previous studies [42]. The longer the course of the disease, the higher the incidence of MACE. In addition, the seasonal onset is a reminder of the importance of being proactive in disease prevention according to the seasonal changes in clinical practice.

The secondary prevention of coronary heart disease consists of two main measures: (1) identification and control of risk factors and (2) appropriate drug therapy [43]. For drug therapy, the use of antiplatelet and anticoagulant drugs after coronary revascularization can reduce the incidence of cardiovascular events [44]. The Chinese expert consensus on the clinical application of perioperative nonoral anticoagulants for PCI published in 2018 [45] states that the perioperative period (before, during, and after PCI) is associated with a high incidence of thrombotic events and therefore anticoagulant

treatment is important. In this study, the incidence of MACE was lower in patients who were taking anticoagulants than for those without anticoagulants, which was consistent with the findings of Song et al [46] showing that routine anticoagulation treatment after surgery may help to reduce the risk of MACE occurrence. However, anticoagulants can also increase the risk of bleeding [47], especially upper gastrointestinal (UGI) bleeding. It is now generally accepted that proton pump inhibitors have a significant protective effect against UGI bleeding caused by antiplatelet and anticoagulant drugs, with omeprazole and lansoprazole being the most potent inhibitors of CYP2C19 [48], which partly explains the lower incidence of MACE in patients using lansoprazole in our study. Additionally, the univariate analysis showed that the incidence of MACE in patients taking diuretics was 8.9% higher than that of patients who were not taking diuretics, which is consistent with previous research [49,50], and is likely related to the fact that diuretics reduce renal blood flow and increase blood concentrations. Therefore, the use of diuretics should be cautiously considered with a full assessment of the fluid status of patients.

Coronary heart disease belongs to the category of “chest pain” or “heartache” in TCM. The main purpose of treatment is to reduce the incidence of angina pectoris, heart failure, myocardial infarction, and other adverse cardiovascular events [51]. In recent years, clinical practice and related studies have confirmed that TCM treatment has some advantages in relieving angina pectoris, intervening restenosis after PCI, preventing and controlling coronary no-reflow after reperfusion, improving quality of life, increasing exercise tolerance, and reducing the incidence of cardiovascular events and adverse reactions [52,53]. Similarly, we found that the variable of TCM treatment was an important feature in the constructed XGBoost model. Therefore, a combination of TCM and western medicine should be considered to provide more beneficial treatment for MACE prevention in practical clinical decision-making, thereby improving the prevention of MACE after coronary revascularization.

With establishment of the bio-psycho-social medical model, the important role of psychological factors on the occurrence and development of diseases is becoming more widely recognized [54]. A large number of evidence-based medical studies have demonstrated the strong relationship between psychological status and the risk of diseases. Barth et al [55] and Roest et al [56] found that depression and anxiety were important risk factors for morbidity and mortality of patients with coronary heart disease, and Taylor et al [57] suggested that depression, social isolation, and emotional abnormalities were closely associated with the occurrence of cardiovascular disease. Patients with coronary revascularization are more likely to suffer from depression and anxiety due to the dual psychological stress of surgery and underlying diseases, and these adverse psychological responses will directly affect prognosis and eventually become risk factors of MACE. For example, by following up 817 patients undergoing CABG for 5.2 years, Blumenthal et al [58] detected that the mortality of patients with moderate to severe depression was 2 to 3 times higher than that of others within 6 months after surgery. Consistent with these findings, we observed that patients with MACE after coronary revascularization had higher HAMA and HAMD scores, indicating greater levels of anxiety and depression. Consequently, it is important to pay more attention to the mental and psychological state of postoperative patients and provide timely psychological guidance and comfort as needed.

Limitations

There are several practical deficiencies and limitations of this study. First, the amount of data available for analysis was limited. It is well known that the performance of ML algorithms depends to a certain extent on the sample size and that the model constructed cannot achieve the best performance, and may even be overfitted, with a small data set. In the future, with data supplementation and further research, we will consider more complex ML algorithms, including deep-learning algorithms, to obtain more accurate and efficient prediction models for clinical observation and research. Second, this was a retrospective study from two centers (ie, The People’s Hospital of Liaoning Province and Affiliated Hospital of Liaoning University of Traditional Chinese Medicine). There is a lack of follow-up data on clinical factors and relevant disease progression; thus, a large multicenter sample study is desired for further generalizability and reliability of the results. Last but not least, in addition to numerical structured data such as vital signs and laboratory tests, clinical electronic medical records also contain a massive amount of unstructured data in the form of text such as patients’ complaints, diagnostic records, and medication information; thus, determining the best ways to use such unstructured information for data analysis and modeling will be the focus of future research. Moreover, we plan to integrate structured and unstructured data comprehensively to develop a risk assessment model to predict the risk probability of MACE in patients with coronary heart disease after revascularization.

Conclusions

In this study, we developed and evaluated risk prediction models for MACE within 6 months after coronary revascularization by utilizing available clinical variables and postoperative follow-up information with ML algorithms. The constructed model can effectively identify high-risk patients with good performance, and the factors that may be associated with MACE were also explored and analyzed in-depth, which is of great significance to provide a reference for medical staff to carry out risk management.

Acknowledgments

This study was supported by the National Science Foundation of China (82174276), China Postdoctoral Foundation (2021M701674), Postdoctoral Research Program of Jiangsu Province (2021K457C), and Qinglan Project of Jiangsu Universities. We thank Jiangsu Famous Medical Technology Co Ltd for providing the data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Table 1. All variables in univariate analysis.
[DOCX File , 35 KB - medinform_v10i4e33395_app1.docx]

References

1. Engelter ST, Reichhart M, Sekoranja L, Georgiadis D, Baumann A, Weder B, et al. Thrombolysis in stroke patients aged 80 years and older: Swiss survey of IV thrombolysis. Neurology 2005 Dec 13;65(11):1795-1798. [doi: 10.1212/01.wnl.0000183702.04080.27] [Medline: 16221951]
2. Sedlis SP, Hartigan PM, Teo KK, Maron DJ, Spertus JA, Mancini GB, COURAGE Trial Investigators. Effect of PCI on long-term survival in patients with stable ischemic heart disease. N Engl J Med 2015 Nov 12;373(20):1937-1946 [FREE Full text] [doi: 10.1056/NEJMoa1505532] [Medline: 26559572]
3. Stone GW, Kappetein AP, Sabik JF, Pocock SJ, Morice M, Puskas J, EXCEL Trial Investigators. Five-year outcomes after PCI or CABG for left main coronary disease. N Engl J Med 2019 Nov 07;381(19):1820-1830. [doi: 10.1056/NEJMoa1909406] [Medline: 31562798]
4. Zhou Y, Zhu R, Chen X, Xu X, Wang Q, Jiang L, et al. Machine learning-based cardiovascular event prediction for percutaneous coronary intervention. J Am Coll Cardiol 2019 Mar;73(9):127. [doi: 10.1016/s0735-1097(19)30735-1]
5. Madhavan MV, Kirtane AJ, Redfors B, Généreux P, Ben-Yehuda O, Palmerini T, et al. Stent-related adverse events >1 year after percutaneous coronary intervention. J Am Coll Cardiol 2020 Feb 18;75(6):590-604 [FREE Full text] [doi: 10.1016/j.jacc.2019.11.058] [Medline: 32057373]
6. Chen X, Liu P, Sun Y, Shen X, Zhang L, Wang X, et al. Research on disease prediction models based on imbalanced medical data sets. Chinese J Comput 2019;42(03):596-609. [doi: 10.11897/SP.J.1016.2019.00596]
7. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975 Aug;8(4):303-320. [doi: 10.1016/0010-4809(75)90009-9] [Medline: 1157471]
8. Dorey FJ. Statistics in brief: Statistical power: what is it and when should it be used? Clin Orthop Relat Res 2011 Feb;469(2):619-620 [FREE Full text] [doi: 10.1007/s11999-010-1435-0] [Medline: 20585913]
9. Kim N, Fischer AH, Dyring-Andersen B, Rosner B, Okoye GA. Research techniques made simple: choosing appropriate statistical methods for clinical research. J Invest Dermatol 2017 Oct;137(10):e173-e178 [FREE Full text] [doi: 10.1016/j.jid.2017.08.007] [Medline: 28941476]
10. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
11. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17 [FREE Full text] [doi: 10.1016/j.csbj.2014.11.005] [Medline: 25750696]
12. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019 Dec 21;19(1):281 [FREE Full text] [doi: 10.1186/s12911-019-1004-8] [Medline: 31864346]
13. Zhu J, Zheng J, Li L, Huang R, Ren H, Wang D, et al. Application of machine learning algorithms to predict central lymph node metastasis in T1-T2, non-invasive, and clinically node negative papillary thyroid carcinoma. Front Med 2021;8:635771. [doi: 10.3389/fmed.2021.635771] [Medline: 33768105]
14. Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease 2015;7(1):129-137. [doi: 10.090592/IJCSC.2016.018]
15. Duan H, Sun Z, Dong W, Huang Z. Utilizing dynamic treatment information for MACE prediction of acute coronary syndrome. BMC Med Inform Decis Mak 2019 Jan 09;19(1):5 [FREE Full text] [doi: 10.1186/s12911-018-0730-7] [Medline: 30626381]
16. Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med 2019 Nov;101:101723. [doi: 10.1016/j.artmed.2019.101723] [Medline: 31813482]
17. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf Sci 2018 Oct;465:1-20. [doi: 10.1016/j.ins.2018.06.056]

18. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Machine Learn Comput 2013 Apr;3(2):224-228. [doi: 10.7763/ijmlc.2013.v3.307]
19. Zhang J, Chen L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. Comput Assist Surg 2019 Oct;24(sup2):62-72. [doi: 10.1080/24699322.2019.1649074] [Medline: 31403330]
20. 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 Jun 04;8(6):e16678 [FREE Full text] [doi: 10.2196/16678] [Medline: 32442149]
21. Ali H, Salleh MNM, Saedudin R, Hussain K, Mushtaq MF. Imbalance class problems in data mining: a review. Indones J Elect Eng Comput Sci 2019 Jun;14(3):1560-1571. [doi: 10.11591/ijeecs.v14.i3.pp1560-1571]
22. Leevy J, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data 2018 Nov 1;5(1):42. [doi: 10.1186/s40537-018-0151-6]
23. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
24. Sun J, Li H, Fujita H, Fu B, Ai W. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. Inf Fusion 2020 Feb;54:128-144. [doi: 10.1016/j.inffus.2019.07.006]
25. Feng S, Keung J, Yu X, Xiao Y, Zhang M. Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction. Inf Soft Technol 2021 Nov;139:106662. [doi: 10.1016/j.infsof.2021.106662]
26. Wang K, Makond B, Chen K, Wang K. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. Appl Soft Comput 2014 Jul;20:15-24. [doi: 10.1016/j.asoc.2013.09.014]
27. Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V, et al. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. IEEE Access 2021;9:39707-39716. [doi: 10.1109/access.2021.3064084]
28. Prusty MR, Jayanthi T, Velusamy K. Weighted-SMOTE: a modification to SMOTE for event classification in sodium cooled fast reactors. Prog Nuclear Energy 2017 Sep;100:355-364. [doi: 10.1016/j.pnucene.2017.07.015]
29. Han H, Wang W, Mao B. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang DS, Zhang XP, Huang GB, editors. Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Berlin, Heidelberg: Springer; 2005:878-887.
30. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. J Healthc Eng 2018;2018:4302425. [doi: 10.1155/2018/4302425] [Medline: 29849998]
31. van Buuren S. Flexible imputation of missing data. New York: Chapman and Hall/CRC; 2012.
32. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Comput Surv 2016 Nov 11;49(2):1-50. [doi: 10.1145/2907070]
33. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 2018 Apr 20;61:863-905. [doi: 10.1613/jair.1.11192]
34. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. Int J Data Mining Knowl Manag Process 2015 Mar 31;5(2):1-11. [doi: 10.5121/ijdkp.2015.5201]
35. Fatourechi M, Ward RK, Mason SG, Huggins J, Schlögl A, Birch GE. Comparison of evaluation metrics in classification applications with imbalanced datasets. 2008 Dec 22 Presented at: 2008 Seventh International Conference on Machine Learning and Applications; December 11-13, 2008; San Diego, CA p. 777-782. [doi: 10.1109/icmla.2008.34]
36. Sun Z, Song Q, Zhu X, Sun H, Xu B, Zhou Y. A novel ensemble method for classifying imbalanced data. Pattern Recogn 2015 May;48(5):1623-1637. [doi: 10.1016/j.patcog.2014.11.014]
37. Hawn MT, Graham LA, Richman JS, Itani KMF, Henderson WG, Maddox TM. Risk of major adverse cardiac events following noncardiac surgery in patients with coronary stents. JAMA 2013 Oct 09;310(14):1462-1472. [doi: 10.1001/jama.2013.278787] [Medline: 24101118]
38. Wu A, Liu J, Zhao Y, Khan AH, Liao L, Tian X. Genetics of coronary artery disease. Sci Sin Vitae 2021 Aug 10;52(2):123-137. [doi: 10.1360/ssv-2020-0347]
39. Wong Y, Cheung CY, Tang CS, Au K, Hai JS, Lee C, et al. Age-biomarkers-clinical risk factors for prediction of cardiovascular events in patients with coronary artery disease. Arterioscler Thromb Vasc Biol 2018 Oct;38(10):2519-2527. [doi: 10.1161/ATVBAHA.118.311726] [Medline: 30354221]
40. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, American Heart Association Council on Epidemiology Prevention Statistics Committee Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2019 Update: a report from the American Heart Association. Circulation 2019 Mar 05;139(10):e56-e528 [FREE Full text] [doi: 10.1161/CIR.0000000000000659] [Medline: 30700139]
41. Ritchie M, Davis J, Aschard H, Battle A, Conti D, Du M, et al. Incorporation of biological knowledge into the study of gene-environment interactions. Am J Epidemiol 2017 Oct 01;186(7):771-777 [FREE Full text] [doi: 10.1093/aje/kwx229] [Medline: 28978191]
42. Luo S, Li Y, Zhao L, He X, Li X, Wang Q, et al. Distribution of traditional Chinese medicine syndromes and relevant factors in 6 months after percutaneous coronary intervention. Chinese J Exp Trad Med Formulae 2020;26(11):194-199. [doi: 10.13422/j.cnki.syfjx.20201122]

43. EUROASPIRE Study Group. EUROASPIRE. A European Society of Cardiology survey of secondary prevention of coronary heart disease: principal results. EUROASPIRE Study Group. European Action on Secondary Prevention through Intervention to Reduce Events. Eur Heart J 1997 Oct;18(10):1569-1582. [doi: 10.1093/oxfordjournals.eurheartj.a015136] [Medline: 9347267]
44. Roffi M, Patrono C, Collet J, Mueller C, Valgimigli M, Andreotti F, et al. 2015 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation. Kardiol Pol 2015 Dec 29;73(12):1207-1294. [doi: 10.5603/kp.2015.0243]
45. Section of Interventional Cardiology of Chinese Society of Cardiology of Chinese Medical Association, Specialty Committee on PreventionTreatment of Thrombosis of Chinese College of Cardiovascular Physicians. Chinese expert consensus on use of parenteral anticoagulants during the peri-percutaneous coronary intervention period. Zhonghua Xin Xue Guan Bing Za Zhi 2018 Jun 24;46(6):428-437. [doi: 10.3760/cma.j.issn.0253-3758.2018.06.005] [Medline: 29925178]
46. Song Y, Tang X, Xu J, Wang H, Liu R, Jiang P, et al. Impact of short-time anticoagulant therapy after selective percutaneous intervention on prognosis of patients with coronary artery disease. Zhonghua Xin Xue Guan Bing Za Zhi 2019 Feb 24;47(2):108-116. [doi: 10.3760/cma.j.issn.0253-3758.2019.02.007] [Medline: 30818938]
47. Feng G, Thanh DN, Jiang X. Predictive factors for postoperative bleeding after percutaneous coronary intervention. Medical Recapitulate 2019;25(13):2611-2616. [doi: 10.3969/j.issn.1006-2084.2019.13.023]
48. Li X, Andersson TB, Ahlström M, Weidolf L. Comparison of inhibitory effects of the proton pump-inhibiting drugs omeprazole, esomeprazole, lansoprazole, pantoprazole, and rabeprazole on human cytochrome P450 activities. Drug Metab Dispos 2004 Aug;32(8):821-827. [doi: 10.1124/dmd.32.8.821] [Medline: 15258107]
49. Nisula S, Kaukonen K, Vaara ST, Korhonen A, Poukkanen M, Karlsson S, FINNAKI Study Group. Incidence, risk factors and 90-day mortality of patients with acute kidney injury in Finnish intensive care units: the FINNAKI study. Intensive Care Med 2013 Mar;39(3):420-428. [doi: 10.1007/s00134-012-2796-5] [Medline: 23291734]
50. Yuan Y, Qiu H, Hu X, Luo T, Gao X, Zhao X, et al. Risk factors of contrast-induced acute kidney injury in patients undergoing emergency percutaneous coronary intervention. Chin Med J 2017;130(1):45-50 [FREE Full text] [doi: 10.4103/0366-6999.196578] [Medline: 28051022]
51. Bi Y, Mao J, Wang X, Zhao Z, Li B, Hou Y. Clinical advantages of traditional Chinese medicine in the prevention and treatment of coronary heart disease and the evaluation of its efficacy. J Trad Chinese Med 2015;56(05):437-440. [doi: 10.13288/j.11-2166/r.2015.05.021]
52. Chen R, Xiao Y, Chen M, He J, Huang M, Hong X, et al. A traditional Chinese medicine therapy for coronary heart disease after percutaneous coronary intervention: a meta-analysis of randomized, double-blind, placebo-controlled trials. Biosci Rep 2018 Oct 31;38(5):BSR20180973 [FREE Full text] [doi: 10.1042/BSR20180973] [Medline: 30143584]
53. Yang J, Tian S, Zhao J, Zhang W. Exploring the mechanism of TCM formulae in the treatment of different types of coronary heart disease by network pharmacology and machining learning. Pharmacol Res 2020 Sep;159:105034. [doi: 10.1016/j.phrs.2020.105034] [Medline: 32565312]
54. Rugulies R. Depression as a predictor for coronary heart disease. a review and meta-analysis. Am J Prev Med 2002 Jul;23(1):51-61. [doi: 10.1016/s0749-3797(02)00439-7] [Medline: 12093424]
55. Barth J, Schumacher M, Herrmann-Lingen C. Depression as a risk factor for mortality in patients with coronary heart disease: a meta-analysis. Psychosom Med 2004;66(6):802-813. [doi: 10.1097/01.psy.0000146332.53619.b2] [Medline: 15564343]
56. Roest AM, Martens EJ, de Jonge P, Denollet J. Anxiety and risk of incident coronary heart disease: a meta-analysis. J Am Coll Cardiol 2010 Jun 29;56(1):38-46 [FREE Full text] [doi: 10.1016/j.jacc.2010.03.034] [Medline: 20620715]
57. Taylor CB, Youngblood ME, Catellier D, Veith RC, Carney RM, Burg MM, ENRICHD Investigators. Effects of antidepressant medication on morbidity and mortality in depressed patients after myocardial infarction. Arch Gen Psychiatry 2005 Jul;62(7):792-798. [doi: 10.1001/archpsyc.62.7.792] [Medline: 15997021]
58. Blumenthal JA, Lett HS, Babyak MA, White W, Smith PK, Mark DB, NORG Investigators. Depression as a risk factor for mortality after coronary artery bypass surgery. Lancet 2003 Aug 23;362(9384):604-609. [doi: 10.1016/S0140-6736(03)14190-6] [Medline: 12944059]

Abbreviations

AUC: area under the curve
CABG: coronary artery bypass grafting
DT: decision tree
FN: false negative
FP: false positive
HAMA: Hamilton anxiety scale
HAMD: Hamilton depression scale
IABP: intraaortic balloon pump
LAD: left atrial diameter

LR: logistic regression
LVEF: left ventricular ejection fraction
MACE: major adverse cardiovascular events
ML: machine learning
NB: naïve Bayes
PCI: percutaneous coronary intervention
RF: random forest
ROC: receiver operating characteristic
SMOTE: synthetic minority oversampling technique
SVM: support vector machine
TCM: traditional Chinese medicine
TN: true negative
TP: true positive
UGI: upper gastrointestinal
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 06.09.21; peer-reviewed by T Kahlon, P Zhao, R Bajpai, JA Benítez-Andrades; comments to author 02.01.22; revised version received 25.02.22; accepted 02.03.22; published 20.04.22.

Please cite as:
Wang J, Wang S, Zhu MX, Yang T, Yin Q, Hou Y
Risk Prediction of Major Adverse Cardiovascular Events Occurrence Within 6 Months After Coronary Revascularization: Machine Learning Study
JMIR Med Inform 2022;10(4):e33395
URL: https://medinform.jmir.org/2022/4/e33395
doi:10.2196/33395
PMID:35442202

Original Paper

Predicting COVID-19 Symptoms From Free Text in Medical Records Using Artificial Intelligence: Feasibility Study

Josefien Van Olmen[1], MD, PhD; Jens Van Nooten[2], MSc; Hilde Philips[1], MD, PhD; Annet Sollie[1], MD, PhD; Walter Daelemans[2], MSc, PhD

[1]Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium

[2]Computational Linguistics, Psycholinguistics and Sociolinguistics Research Centre, University of Antwerp, Antwerp, Belgium

Corresponding Author:
Josefien Van Olmen, MD, PhD
Department of Family Medicine and Population Health
University of Antwerp
Prinsstraat 13
Antwerp, 2000
Belgium
Phone: 32 475892225
Email: josefien.vanolmen@uantwerpen.be

Abstract

Background: Electronic medical records have opened opportunities to analyze clinical practice at large scale. Structured registries and coding procedures such as the International Classification of Primary Care further improved these procedures. However, a large part of the information about the state of patient and the doctors' observations is still entered in free text fields. The main function of those fields is to report the doctor's line of thought, to remind oneself and his or her colleagues on follow-up actions, and to be accountable for clinical decisions. These fields contain rich information that can be complementary to that in coded fields, and until now, they have been hardly used for analysis.

Objective: This study aims to develop a prediction model to convert the free text information on COVID-19–related symptoms from out of hours care electronic medical records into usable symptom-based data that can be analyzed at large scale.

Methods: The design was a feasibility study in which we examined the content of the raw data, steps and methods for modelling, as well as the precision and accuracy of the models. A data prediction model for 27 preidentified COVID-19–relevant symptoms was developed for a data set derived from the database of primary-care out-of-hours consultations in Flanders. A multiclass, multilabel categorization classifier was developed. We tested two approaches, which were (1) a classical machine learning–based text categorization approach, Binary Relevance, and (2) a deep neural network learning approach with BERTje, including a domain-adapted version. Ethical approval was acquired through the Institutional Review Board of the Institute of Tropical Medicine and the ethics committee of the University Hospital of Antwerpen (ref 20/50/693).

Results: The sample set comprised 3957 fields. After cleaning, 2313 could be used for the experiments. Of the 2313 fields, 85% (n=1966) were used to train the model, and 15% (n=347) for testing. The normal BERTje model performed the best on the data. It reached a weighted F1 score of 0.70 and an exact match ratio or accuracy score of 0.38, indicating the instances for which the model has identified all correct codes. The other models achieved respectable results as well, ranging from 0.59 to 0.70 weighted F1. The Binary Relevance method performed the best on the data without a frequency threshold. As for the individual codes, the domain-adapted version of BERTje performs better on several of the less common objective codes, while BERTje reaches higher F1 scores for the least common labels especially, and for most other codes in general.

Conclusions: The artificial intelligence model BERTje can reliably predict COVID-19–related information from medical records using text mining from the free text fields generated in primary care settings. This feasibility study invites researchers to examine further possibilities to use primary care routine data.

(JMIR Med Inform 2022;10(4):e37771) doi:10.2196/37771

KEYWORDS

natural language processing; text mining; electronic medical records; COVID-19; structured registry; coding procedure; prediction model; feasibility study; precision model; artificial intelligence; primary care

Introduction

Electronic medical records (EMRs) have opened the opportunity to analyze clinical practice at large scale, and to perform clinical-epidemiological research, which can inform health care managers and policy makers. Structured registries and coding procedures such as the International Classification of Primary Care have improved the way doctors put information into EMR, which has facilitated the use of its output and accelerated research using these data. The free text fields also still available in EMR systems have been hardly used apart from clinical follow-up. Yet the usage of this information has great potential to contribute to monitoring and evaluation of clinical practice and to EMR-driven research. In 2016, US researchers compared the accuracy for case detection of diagnoses such as dementia, stroke, diabetes, and depression based upon coded information versus the procedure including free text, and they found a significant improvement in algorithm sensitivity in the latter [1].

This is not surprising since these fields contain the core of clinical practice captured in the encounter notes. The encounter notes available in most EMRs have a structured "SOAP" format, which stands for Subjective (patient's history), Objective (physical examination), Assessment (initial differential diagnosis), and Plan [2]. The main function of these free text fields is to report the doctor's line of thought, to remind oneself and colleagues on follow-up actions, and to be accountable for clinical decisions. Therefore, they contain the richest data about the state of the patient and the observations of the doctor. Yet their use is also challenging. Health care providers tend to write notes quickly, with personal styles and abbreviations, and they vary in their completeness and quality of reporting. Therefore, encounter notes have seldom been used for further analyses and research.

A 2019 review on the use of free text fields in the EMR [3] showed that the focus of most studies was on the development of methods to extract symptom information for disease classification tasks. For instance, a UK study validated a method for mining free text fields to link them to frequent medical conditions such as colic or renal failure [4]. The analysis of symptoms themselves has been restricted to specific and rather narrow domains such as neuromuscular diseases [5], psychiatry [6], and veterinary medicine [7,8]. A recent study demonstrates the feasibility of extracting information from free text notes and using this as input to a model for predicting patient outcomes [9].

To use the information from free text fields at a large scale, methods to recognize this information need to be developed and evaluated. A 2012 study found that combination of a manually created filter and rule learning algorithm yielded the best performance across two different data sets (radiology reports and general practitioner [GP] notes) [10], but the performance for the GP set was considerably lower. The variation of symptoms and note-taking is peculiar for the GP domain. This implies that more such studies are necessary to develop robust methods for data recognition for GP data sets to improve the reproducibility of data and their value for routine use.

The relevance for quick information using real time data was apparent in the COVID-19 pandemic. The collection, evaluation, and synthesis of information started quickly. Data mainly came from hospital settings, where most severe cases were admitted, and where resources could be mobilized quickly, for instance, to make decision-support algorithms for diagnosis and treatment based upon models that predict disease outcomes [11]. This predominant use of data from severely ill patients led to risk of bias in the models [12]. This underlined the need to develop methods to extract data quickly and reliably from primary care health records at large scale.

Our study contributes to this goal. The objective of this paper was to develop a robust method to transform the primary care notes into a list of symptoms that could feed improved COVID-19 prediction models through the development of a text classifier model that can predict the relevant symptoms (output) based upon the analysis of the free text fields (input). If this method proves robust, free text data from primary care clinical notes about COVID-19–related symptoms can be mined at large scale quickly and reliably.

Methods

Background

This study is part of the project ID-CoV to develop procedures for data identification, harmonization, and linkage to develop robust methodologies to build a risk prediction tool based on primary care and hospital data for the identification of individuals at higher risk for severe COVID-19 outcomes (project id 43639, Funded by University of Antwerp).

Data Collection

The iCAREdata database was used, which is a database of contacts in out of hours (OOH) care by general practice cooperatives, triage centers (additional centers organized during the COVID-19 pandemic to triage between infectious and noninfectious diseases), pharmacies, and a small number of first aid departments connected to the system (covering OOH care of roughly two-thirds of Flanders population) [13]. One OOH hosts between 80 and 150 different GPs. Data from EMR at OOH services therefore cover a broad range of different physicians, with different approaches of medical care and registration of clinical data, leading to high variability of content, completeness, quality, and format of information in the data set, which adds methodological challenges to developing mining procedures. Nevertheless, the analysis of the data of this segment of primary care consultations is especially relevant in a pandemic context [14]. The units of analysis in iCAREdata are records, each record being one contact (=consultation). Due to the exploratory nature, sample size was not considered a limiting factor. We aimed to use as many observations (patient's encounters) as possible in a given time period to reduce the uncertainty of our model estimates. A study database was created that comprises all records from January 1, 2019, to November 30, 2020. These are roughly 779,000 records, which

include a pre–COVID-19 period and a COVID-19 epidemic period (March 1, 2020, to November 30, 2020).

For each record, 15 fields were extracted (Multimedia Appendix 1). For the data mining study reported in this paper, only 5 fields were used (Textbox 1). The "field subjective" (physician's report on the patient's account of their problem) and "field objective" (findings and measurements of the physician) were explored for relevant text (combinations). We used supervised machine learning algorithms to classify information into one or more of predetermined symptoms via the multiclass, multilabel prediction model described below. Fields "DiagnTekst" and "DiagnCod" were used as control records for validation.

The establishment of the symptom list that needed to be the outcome of the classifier model was started from an initial list of 23 symptoms identified by the Belgium Public Health Institute as relevant [15] but was refined driven by the data. A manual exploration of the data set yielded 62 symptoms most of them with a negative counterpart, indicating the absence of that symptom. Negative symptoms were relevant because of their negative predictive value in a diagnostic or prognostic algorithm [16]; for instance, the absence of cough contributing to the likelihood or non-likelihood of a COVID-19 diagnosis. The skewed distribution led to a regrouping of symptoms, resulting in a final list of 27 signs or symptoms (Table 1). There are two types of symptom codes, which are "objective," based on the "objectief" text field, and "subjective," based on the "subjectief" text field, respectively.

Textbox 1. Relevant fields for input to machine learning algorithm to recognize signs and symptoms.

Machine learning fields

- IdContact: unique id for contact (date, guard post, time)
- Subjectief: subjective text field
- Objectief: objective text field
- DiagnTekst: diagnosis term (thesaurus)
- DiagnCod: diagnosis code from the International Classification of Primary Care [17]

Table 1. Final list with signs and symptoms to be coded from the free text.

Final symptoms—coded	Explanation
S[a]1; SA[b]1	Cough
S100; SA100	Upper respiratory tract infection complaints
S101; SA101	Dyspnea and shortness of breath
S7; SA7	Thoracic pain or chest pain
S102; SA102	Loss of taste or smell
S10; SA10	History of fever
S112	Pain or stiffness in muscles, joints, or neck
S109	Complaints of throat or voice
S12	Fatigue
S15	Headache
S103; SA103	Gastrointestinal complaints
S104	Significant acute event or change
S105	Chronic pulmonary complaints; smoking; potentially worsening
S105	Other comorbidities or being pregnant
S106	Known cardiovascular diseases or hypertension or relevant medication
S107	Known diabetes or diabetes medication
S108	Medication NSAID[c] or immunosuppressive drugs
S113	Palpitations or dizziness
S110	General complaints as malaise and illness
S111	Mental or sleeping problems
S63	Close contact with a sick person (COVID-19 symptoms) or COVID-19–positive case
O[d]101	Respiratory signs found during physical examination
O6	Fever measured by health care staff
O102	Ear-, nose-, or throat-positive signs during physical examination
O104	Neurological symptoms
O103	Circulatory positive signs: abnormal pulse rate, tension, or turgor of capillary refill
O19	Impression of being ill

[a]S: Subjective.

[b]A: absence of the symptom.

[c]NSAID: nonsteroidal anti-inflammatory drugs.

[d]O: Objective.

Development of a Classifier Model

Classification entails the tasks of predicting the class (or label of output variable—the list with 27 signs or symptoms) based upon the input variables (the free text fields). Two approaches were examined to develop a multiclass, multilabel categorization classifier, which are as follows: (1) a classical machine learning–based text categorization approach; and (2) a deep neural network learning approach based on fine-tuning a pretrained model for domain adaptation and learning the classification task. The advantage of the latter approach is that, in general, less supervised training data (ie, annotated data) are needed for learning the task. A random sample from the data set was extracted for annotation, with a distribution of 1/3 records from before the start of the COVID-19 pandemic (operationalized as March 1, 2020) and 2/3 after that date, comprising 3957 entries in total. Character encoding problems in the text data were solved during preprocessing. Empty entries and entries that did not contain any information (eg, "/") in either the subjective or objective fields were removed from the data set, which left 2313 entries to be used for the experiments. The subjective and objective text fields were merged into one text field in order to receive sufficiently large text fragments for prediction. The same resulting text could be assigned multiple objective and subjective codes. Negative symptoms were kept apart by coding them with an A-label; for instance, SA10 indicated the absence of a history of fever. The A codes were frequent among the objective text fields. Entries that were

annotated as irrelevant (without any symptom code) were used as negative examples for training of the models.

The samples were annotated by 5 medical doctors or researchers. Inter-annotator variability was checked. All annotators started annotation of the same set and manually compared inconsistencies, discussed them, adapted the standard operating guidelines, and repeated this procedure until agreement of 90% was achieved. During the annotation phase, the inventory of symptom tags (classes) evolved, but all annotated data were made comparable through a common code book and standard operating procedure in the final data set. The number of entries, average number of tokens (instances of words and punctuation marks), and total amount of tokens for the training partition, test partition, and the total data set are summarized in Table 2.

The distribution of codes (labels) in the data set is shown in Figures 1 and 2. The majority of the codes are subjective codes; out of the 55 codes, 43 (78%) are subjective while the remaining 12 (22%) are objective. For the development of the classifier, experiments were conducted with all codes and only codes occurring at least 50 times, which meant 35 (63%) out of 55 codes (representing 93% of all used codes).

Table 2. Total number of entries, average amount of tokens per entry, and total amount of tokens for the training, test portions, and the entire data set.

Portion	Entries, n (%)	Average tokens per entry, n	Total tokens, n
Train	1966 (85)	24	53,929
Test	347 (15)	31	10,779
Total	2313 (100)	28	64,708

Figure 1. Code distribution in the data set. Codes to the right of the threshold line were removed for the experiments where a frequency threshold was employed.

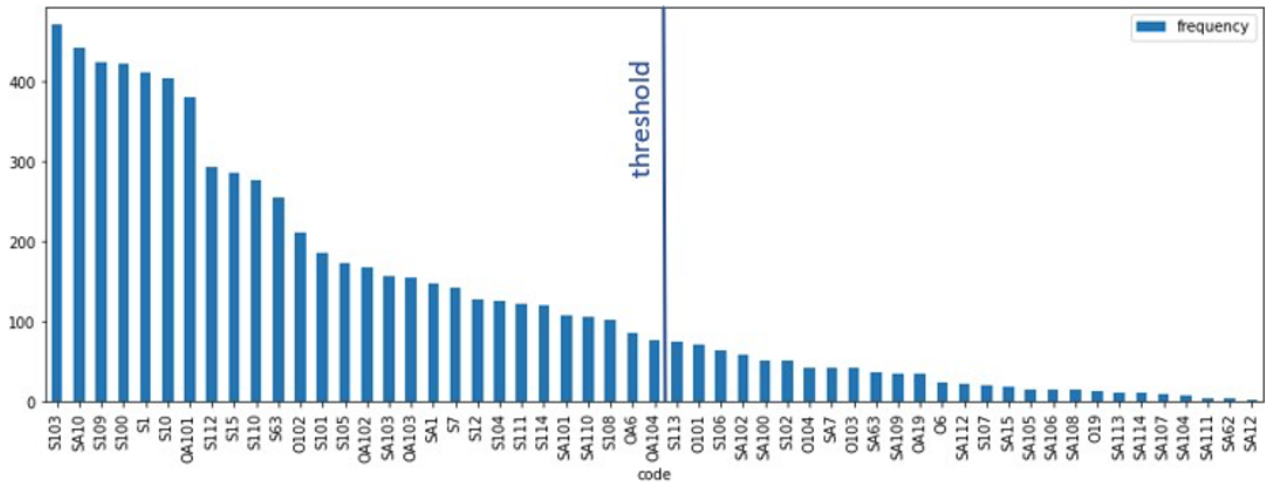
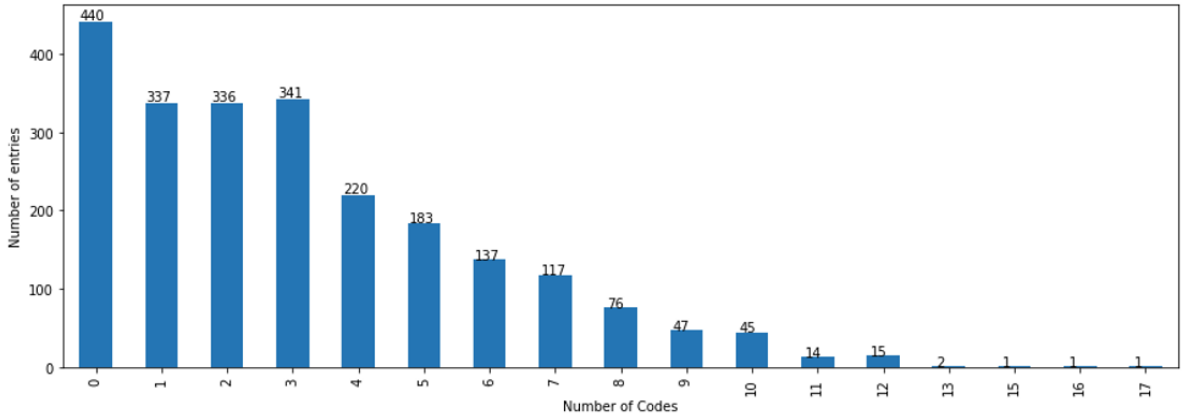


Figure 2. Distribution of the percentage of entries in the data set assigned to a particular number of codes.



The baseline accuracies (most frequent class prediction and random prediction) are 0.15 and 0.08, respectively. In the first set of experiments, we used classic machine learning methods. One of the most common approaches to multiclass, multilabel classification is Binary Relevance. With this method, the multilabel problem is translated to *n* binary classification problems, where *n* is equal to the number of labels present in the data set. Binary in this case means that the classifier attempts to predict whether a class (code) is present (1) or not (0) in the text. For the binary classifiers, we used the Stochastic Gradient

Descent classifier [18] and optimized the hyperparameters (including the loss function) by performing a gridsearch on them (a search for the best combination of algorithm parameters on a validation partition of the training data in the context of 5-fold cross-validation). The performance of this method is measured by taking the mean of all cross-validated results from the individual binary classifiers.

Further experiments were then conducted with BERTje [19], a Dutch version of BERT [20]. BERT is a widely used model for natural language processing, and the availability of a Dutch version BERTje made it the first choice of the team. BERTje is an open-source pretrained language model that has been trained on a large amount of generic (nonmedical) Dutch text data. Thus, the model already has knowledge about language patterns before having been trained on data for a specific problem, in contrast to, for example, the Stochastic Gradient Descent classifier, which was limited to the training data. Additionally, we continued the pretraining of BERTje by using a selection of the text fields of the original data set (part of the iCAREdata database) in order to "adapt" BERTje to medical texts. This method has been proven to be successful on a wide range of tasks [21,22]. For all experiments, the F1 macro score metric was used for evaluation, which is the average F1 score (harmonic mean of precision and recall) obtained for the classes. In our binary relevance setup and the implementation of F1 macro we used, only successful predictions of the minority class (correctly predicting that the code is present) are taken into account, which makes it the most critical (but also the most relevant) evaluation.

For all experiments, we used a stratified train-test split, where 80% of the data were used for training and hyperparameter optimization, and 20% were used for testing. The best model on test (BERTje) was then fine-tuned on all annotated data and applied to the complete (unannotated) data set, predicting diagnostic codes based on the text fields.

Ethics Approval

Ethical approval was acquired through the Institutional Review Board of the Institute of Tropical Medicine and the ethics committee of the University Hospital of Antwerpen (ref 20/50/693).

Results

In the tables below, the results of the experiments on the test set are summarized. Across all models that were trained and tested on data with a frequency threshold for the labels, the normal BERTje model performed the best on the data, reaching a weighted F1 score of 0.70 and an exact match ratio or accuracy score of 0.38 (Table 3), indicating the instances for which the model has identified all correct codes. The results per code can be found in Table S1 of Multimedia Appendix 1. The other models achieved respectable results as well, ranging from 0.59 to 0.70 weighted F1. The Binary Relevance method performed the best on the data without a frequency threshold (Table S2 of Multimedia Appendix 1).

Regarding the results on the individual codes themselves, the domain-adapted version of BERTje performs better on several of the less common objective codes (O101, O102, OA101, OA102, OA104, and OA6), while BERTje reaches higher F1 scores for the least common labels (S102 and SA102) especially, and most other codes in general.

Table 3. Average results for the different models on test data with a frequency threshold for the codes (codes occurring at least 50 times).

Method	Weighted precision	Weighted specificity	Weighted recall	Weighted F1
Binary Relevance (SGD[a] classifier)	0.69	0.93	0.52	0.59
BERTje	0.77	0.97	0.68	0.70
BERTje (domain adaptation)	0.74	0.96	0.62	0.67

[a]SGD: Stochastic Gradient Descent.

Discussion

Principal Findings

In this paper, we demonstrated the feasibility of developing a model to predict symptom codes from primary care clinical text notes. Across the three models tested, the pretrained neural network model BERTje performed the best. The reason for the lower performance of the domain-adapted BERTje needs further investigation. Neural networks can forget information they previously learned upon learning new information (catastrophic forgetting); however, from the current data, we are not able to explain if this was the reason for the lower performance.

Our model resulted in the ability to predict symptoms from the free text with a weighted average F score of 0.66 (0.75 sensitivity and 0.97 specificity) on all codes, regardless of frequency, and an F score of 0.70 (0.77 sensitivity and 0.97 specificity) on codes that occurred more than 50 times in the data set. Very few studies that have developed mining techniques for clinical notes, in general [23], and from primary care, in particular. Yet the incidental other studies show feasibility and good results [24]. A study using a Repeated Incremental Pruning to Produce Error Reduction rule learning model resulted in a sensitivity of 0.91, and a specificity 0.76 [10]. To our knowledge, this is the first study that mined data from OOH health care organizations.

The strength of our study is that we used a large database representative of a population of 6 million people in Flanders and with many different GPs. The major limitation of our study relates to the quality of the raw data. The data set contained consultations of OOH primary care consultations. The notes in these consultations were often very brief, and the completeness and quality of information varied across entries. This is similar in studies from routine primary care [25]; however, in OOH

care, this is likely to be worse, making it more difficult to develop mining models. This reflects the reality of medical practice and the limitations of real-world data. Further research into minimal needs for reporting for both clinical and other purposes is warranted. Another limitation is that some symptom codes, for instance SA100 (*geen BLWI klachten-no respiratory tract complaints)* could not be learned by the machine learning models. The explanation for this, as for similar cases, is that there were too few instances available in the data set for the model to learn from [9]. For these codes, it would be useful to investigate the data for more cases to be annotated. Even if more elaborate annotating will improve the gain, not all free text fields can be transformed into coded information, which needs to be taken into account in the interpretation of the output.

Notwithstanding the limitations, our study is relevant for primary care research and evaluation. Once coded, these symptoms can be monitored, evaluated, and processed, for the development and testing of algorithms, for near real time symptom surveillance [26], or for assessing quality of history taking and record keeping. Our study focused on symptom detection, but wider applications of the text mining and natural language processing can be thought of, such as the analyses of adverse events or patient-reported experiences [23].

Conclusions

The BERTje prediction models can reliably predicting COVID-19–related information from medical records using text mining from the free text fields generated in primary care settings. The feasibility to convert this rich but largely untapped source of clinical encounter into data usable for monitoring, evaluation, and research provides opportunities for comprehensive analysis of primary care consultations at large scale, as well as use for monitoring purposes, also in other primary care settings. This feasibility study invites researchers to examine further possibilities to use primary care routine data, for instance, to examine the process of clinical reasoning through EMR analysis or to assess the input of patient-related information into the diagnostic process.

Acknowledgments

We acknowledge the following people: Nathalie Wisse and Veronique Verhoeven for their contribution to the manual coding process; José Peñalvo, Elly Mertens, and Els van Gentbrugge for the development of the jointly funded project ID-COV; and the iCAREdata team for extracting the data set. This study was funded by the University of Antwerp (ID 43639) through a joint Pump Priming Proposal Fund.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details about experiments.
[DOCX File , 39 KB - medinform_v10i4e37771_app1.docx]

References

1. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep 05;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
2. Pearce PF, Ferguson LA, George GS, Langford CA. The essential SOAP note in an EHR age. Nurse Pract 2016 Feb 18;41(2):29-36. [doi: 10.1097/01.NPR.0000476377.35114.d7] [Medline: 26795838]
3. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
4. Duz M, Marshall JF, Parkin T. Validation of an Improved Computer-Assisted Technique for Mining Free-Text Electronic Medical Records. JMIR Med Inform 2017 Jun 29;5(2):e17 [FREE Full text] [doi: 10.2196/medinform.7123] [Medline: 28663163]
5. Kaya H, Alcan V, Zinnuroğlu M, Karataş GK, Çoban S, Dolgun M, et al. Analysis of free text in electronic health records by using text mining methods. 2018 Presented at: 7th International Conference on Advanced Technologies(ICAT'18); 28 April - 01 May 2018; Antalya, Turkey.
6. Karystianis G, Nevado AJ, Kim C, Dehghan A, Keane JA, Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. Int J Methods Psychiatr Res 2018 Mar 22;27(1):e1602 [FREE Full text] [doi: 10.1002/mpr.1602] [Medline: 29271009]
7. Anholt R, Berezowski J, Jamal I, Ribble C, Stephen C. Mining free-text medical records for companion animal enteric syndrome surveillance. Prev Vet Med 2014 Mar 01;113(4):417-422. [doi: 10.1016/j.prevetmed.2014.01.017] [Medline: 24485708]

8. Welsh CE, Duz M, Parkin TD, Marshall JF. Disease and pharmacologic risk factors for first and subsequent episodes of equine laminitis: A cohort study of free-text electronic medical records. Prev Vet Med 2017 Jan 01;136:11-18. [doi: 10.1016/j.prevetmed.2016.11.012] [Medline: 28010903]
9. Goh KH, Wang L, Yeow AYK, Ding YY, Au LSY, Poh HMN, et al. Prediction of Readmission in Geriatric Patients From Clinical Notes: Retrospective Text Mining Study. J Med Internet Res 2021 Oct 19;23(10):e26486 [FREE Full text] [doi: 10.2196/26486] [Medline: 34665149]
10. Schuemie MJ, Sen E, 't Jong GW, van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. Pharmacoepidemiol Drug Saf 2012 Jun 24;21(6):651-658. [doi: 10.1002/pds.3205] [Medline: 22271492]
11. Jimenez-Solem E, Petersen TS, Hansen C, Hansen C, Lioma C, Igel C, et al. Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. Sci Rep 2021 Feb 05;11(1):3246 [FREE Full text] [doi: 10.1038/s41598-021-81844-x] [Medline: 33547335]
12. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
13. Colliers A, Bartholomeeusen S, Remmen R, Coenen S, Michiels B, Bastiaens H, et al. Improving Care And Research Electronic Data Trust Antwerp (iCAREdata): a research database of linked data on out-of-hours primary care. BMC Res Notes 2016 May 04;9(1):259 [FREE Full text] [doi: 10.1186/s13104-016-2055-x] [Medline: 27142361]
14. Morreel S, Philips H, Verhoeven V. Organisation and characteristics of out-of-hours primary care during a COVID-19 outbreak: A real-time observational study. PLoS One 2020 Aug 13;15(8):e0237629 [FREE Full text] [doi: 10.1371/journal.pone.0237629] [Medline: 32790804]
15. Gevalsdefinitie, indicaties voor testen en verplichte melding van covid-19. Sciensano. 2020. URL: https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_Case%20definition_Testing_NL.pdf [accessed 2021-12-06]
16. Tostmann A, Bradley J, Bousema T, Yiek W, Holwerda M, Bleeker-Rovers C, et al. Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. Euro Surveill 2020 Apr;25(16):pii=2000508 [FREE Full text] [doi: 10.2807/1560-7917.ES.2020.25.16.2000508] [Medline: 32347200]
17. ICPC-2 International Classification of Primary Care - 2nd edition. Universiteit Gent. URL: https://www.transhis.nl/wp-content/uploads/2014/12/icpc-2-2pager-nederlands.pdf [accessed 2022-04-22]
18. Robbins H, Monro S. A Stochastic Approximation Method. Ann. Math. Statist 1951 Sep;22(3):400-407. [doi: 10.1214/aoms/1177729586]
19. De Vries W, Van Cranenburgh A, Bisazza A, Caselli T, Van Noord G, Nissim M. BERTje: A Dutch BERT Model. GitHub. URL: https://github.com/cl-tohoku/bert-japanese [accessed 2022-02-23]
20. Devlin J, Chang M, Lee K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. GitHub. URL: https://github.com/tensorflow/tensor2tensor [accessed 2022-02-23]
21. Rietzler A, Stabinger S, Opitz P, Engl S. Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification. arXiv 2020:11-16. [doi: 10.48550/arXiv.1908.11860]
22. Han X, Eisenstein J. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. arXiv 2022:4238-4248 [FREE Full text] [doi: 10.18653/v1/d19-1433]
23. Hendrickx I, Voets T, van Dyk P, Kool RB. Using Text Mining Techniques to Identify Health Care Providers With Patient Safety Problems: Exploratory Study. J Med Internet Res 2021 Jul 27;23(7):e19064 [FREE Full text] [doi: 10.2196/19064] [Medline: 34313604]
24. Hardjojo A, Gunachandran A, Pang L, Abdullah MRB, Wah W, Chong JWC, et al. Validation of a Natural Language Processing Algorithm for Detecting Infectious Disease Symptoms in Primary Care Electronic Medical Records in Singapore. JMIR Med Inform 2018 Jun 11;6(2):e36 [FREE Full text] [doi: 10.2196/medinform.8204] [Medline: 29907560]
25. Seo J, Kong H, Im S, Roh H, Kim D, Bae H, et al. A pilot study on the evaluation of medical student documentation: assessment of SOAP notes. Korean J Med Educ 2016 Jun;28(2):237-241 [FREE Full text] [doi: 10.3946/kjme.2016.26] [Medline: 26996436]
26. Birtwhistle R, Williamson T. Primary care electronic medical records: a new data source for research in Canada. CMAJ 2015 Mar 03;187(4):239-240 [FREE Full text] [doi: 10.1503/cmaj.140473] [Medline: 25421989]

Abbreviations

EMR: electronic medical record
GP: general practitioner
OOH: Out of Hours

Edited by C Lovis; submitted 06.03.22; peer-reviewed by V Tiberius; comments to author 27.03.22; revised version received 31.03.22; accepted 11.04.22; published 27.04.22.

Please cite as:
Van Olmen J, Van Nooten J, Philips H, Sollie A, Daelemans W
Predicting COVID-19 Symptoms From Free Text in Medical Records Using Artificial Intelligence: Feasibility Study
JMIR Med Inform 2022;10(4):e37771
URL: https://medinform.jmir.org/2022/4/e37771
doi:10.2196/37771
PMID:35442903

Corrigenda and Addenda

Metadata Correction: A Comparison of Different Modeling Techniques in Predicting Mortality With the Tilburg Frailty Indicator: Longitudinal Study

Tjeerd van der Ploeg[1], PhD; Robbert Gobbens[1,2,3], PhD

[1]Faculty of Health, Sports and Social Work, Inholland University of Applied Sciences, Amsterdam, Netherlands

[2]Zonnehuisgroep Amstelland, Amstelveen, Netherlands

[3]Department Family Medicine and Population Health, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium

Corresponding Author:
Tjeerd van der Ploeg, PhD
Faculty of Health, Sports and Social Work
Inholland University of Applied Sciences
De Boelelaan 1109
Amsterdam, 1081 HV
Netherlands
Phone: 31 653519264
Email: tvdploeg@quicknet.nl

Related Article:

Correction of: https://medinform.jmir.org/2022/3/e31480

(JMIR Med Inform 2022;10(4):e31479) doi:10.2196/31479

In "A Comparison of Different Modeling Techniques in Predicting Mortality With the Tilburg Frailty Indicator: Longitudinal Study" (JMIR Med Inform 2022;10(3):e31480) the authors noted the following two errors.

1. In the originally published article, author affiliations appeared as follows:

> *Tjeerd van der Ploeg, PhD; Robbert Gobbens, PhD*
>
> *Faculty of Engineering, Design and Computer Technology, Inholland University of Applied Sciences, Alkmaar, Netherlands*

The corrected affiliations for the authors are as follows:

> *Tjeerd van der Ploeg[1], PhD; Robbert Gobbens[1,2,3], PhD*
>
> *[1]Faculty of Health, Sports and Social Work, Inholland University of Applied Sciences, Amsterdam, Netherlands*
>
> *[2]Zonnehuisgroep Amstelland, Amstelveen, Netherlands*
>
> *[3]3Department Family Medicine and Population Health, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium*

2. In the originally published article, the corresponding author's address appeared as follows:

> *Tjeerd van der Ploeg, PhD*
>
> *Faculty of Engineering, Design and Computer Technology*
>
> *Inholland University of Applied Sciences*
>
> *Bergerweg 200*
>
> *Alkmaar, 1817 MR*
>
> *Netherlands*

The address has been corrected as follows:

> *Tjeerd van der Ploeg, PhD*
>
> *Faculty of Health, Sports and Social Work*
>
> *Inholland University of Applied Sciences*
>
> *De Boelelaan 1109*
>
> *Amsterdam, 1081 HV*
>
> *Netherlands*

The correction will appear in the online version of the paper on the JMIR Publications website on April 8, 2022, 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 05.04.22; this is a non–peer-reviewed article;accepted 05.04.22; published 08.04.22.

Please cite as:
van der Ploeg T, Gobbens R
Metadata Correction: A Comparison of Different Modeling Techniques in Predicting Mortality With the Tilburg Frailty Indicator: Longitudinal Study
JMIR Med Inform 2022;10(4):e31479
URL: https://medinform.jmir.org/2022/4/e31479
doi:10.2196/31479
PMID:35394921

Corrigenda and Addenda

Correction: Mining Electronic Health Records for Drugs Associated With 28-day Mortality in COVID-19: Pharmacopoeia-wide Association Study (PharmWAS)

Ivan Lerner[1,2,3], MD; Arnaud Serret-Larmande[1,2], MD; Bastien Rance[1,3], PhD; Nicolas Garcelon[3,4], PhD; Anita Burgun[1,2,3], MD, PhD; Laurent Chouchana[5], PhD, PharmD; Antoine Neuraz[1,2,3], MD, PhD

[1]Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Paris, France

[2]Informatique biomédicale, Hôpital Necker-Enfants Malades, Assistance Publique - Hôpitaux de Paris, Paris, France

[3]HeKA Team, Inria, Paris, France

[4]Inserm UMR 1163, Data Science Platform, Université de Paris, Imagine Institute, Paris, France

[5]Centre Régional de Pharmacovigilance, Service de Pharmacologie, Hôpital Cochin, Assistance Publique - Hôpitaux de Paris, Centre - Université de Paris, Paris, France

Corresponding Author:
Antoine Neuraz, MD, PhD
Inserm
Centre de Recherche des Cordeliers
Sorbonne Université
Université de Paris
15 Rue de l'École de Médecine
Paris, 75006
France
Phone: 33 01 44 27 64 82
Email: antoine.neuraz@aphp.fr

Related Article:

Correction of: https://medinform.jmir.org/2022/3/e35190

(JMIR Med Inform 2022;10(4):e38505) doi:10.2196/38505

In "Mining Electronic Health Records for Drugs Associated With 28-day Mortality in COVID-19: Pharmacopoeia-wide Association Study (PharmWAS)" (JMIR Med Inform 2022;10(3):e35190), the following corrections were made.

1. In the Results section of the abstract, a Q-value was incorrectly written as follows:

> *Among these, diazepam and tramadol were the only ones not discarded by automated diagnostics, with adjusted odds ratios of 2.51 (95% CI 1.52-4.16, Q=.1) and 1.94 (95% CI 1.32-2.85, Q=.02), respectively.*

This has been corrected to:

> *Among these, diazepam and tramadol were the only ones not discarded by automated diagnostics, with adjusted odds ratios of 2.51 (95% CI 1.52-4.16, Q=.01) and 1.94 (95% CI 1.32-2.85, Q=.02), respectively.*

2. In the first paragraph of the discussion, the following sentence was incorrectly added as follows:

> *Indeed, of 87 treatments prescribed in the first 48 hours, 4 (5%) were associated with increased 28-day mortality after adjustment of confounding factors and multiple testing correction, and none were associated with increased mortality.*

This has been corrected to:

> *Indeed, of 87 treatments prescribed in the first 48 hours, 4 (5%) were associated with increased 28-day mortality after adjustment of confounding factors and multiple testing correction, and none were associated with decreased mortality.*

The correction will appear in the online version of the paper on the JMIR Publications website on April 12, 2022, 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 05.04.22; this is a non–peer-reviewed article;accepted 05.04.22; published 12.04.22.

Please cite as:
Lerner I, Serret-Larmande A, Rance B, Garcelon N, Burgun A, Chouchana L, Neuraz A
Correction: Mining Electronic Health Records for Drugs Associated With 28-day Mortality in COVID-19: Pharmacopoeia-wide Association Study (PharmWAS)
JMIR Med Inform 2022;10(4):e38505
URL: https://medinform.jmir.org/2022/4/e38505
doi:10.2196/38505
PMID:35413000

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org