JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 10 (2022), Issue 5 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Review

Computerized Clinical Decision Support Systems for the Early Detection of Sepsis Among Pediatric, Neonatal, and Maternal Inpatients: Scoping Review

Khalia Ackermann[1], MPH; Jannah Baker[1], PhD; Marino Festa[2], MD(Res); Brendan McMullan[3,4], PhD; Johanna Westbrook[1], PhD; Ling Li[1], PhD

[1]Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Macquarie University, Australia

[2]Kids Critical Care Research, Department of Paediatric Intensive Care, Children's Hospital at Westmead, Sydney, Australia

[3]Department of Immunology and Infectious Diseases, Sydney Children's Hospital, Randwick, Sydney, Australia

[4]Faculty of Medicine & Health, University of New South Wales, Sydney, Australia

Corresponding Author:
Khalia Ackermann, MPH
Centre for Health Systems and Safety Research
Australian Institute of Health Innovation
Level 6, 75 Talavera road
Macquarie University, 2109
Australia
Phone: 61 2 9850 2432
Email: khalia.ackermann@mq.edu.au

Abstract

Background: Sepsis is a severe condition associated with extensive morbidity and mortality worldwide. Pediatric, neonatal, and maternal patients represent a considerable proportion of the sepsis burden. Identifying sepsis cases as early as possible is a key pillar of sepsis management and has prompted the development of sepsis identification rules and algorithms that are embedded in computerized clinical decision support (CCDS) systems.

Objective: This scoping review aimed to systematically describe studies reporting on the use and evaluation of CCDS systems for the early detection of pediatric, neonatal, and maternal inpatients at risk of sepsis.

Methods: MEDLINE, Embase, CINAHL, Cochrane, Latin American and Caribbean Health Sciences Literature (LILACS), Scopus, Web of Science, OpenGrey, ClinicalTrials.gov, and ProQuest Dissertations and Theses Global (PQDT) were searched by using a search strategy that incorporated terms for sepsis, clinical decision support, and early detection. Title, abstract, and full-text screening was performed by 2 independent reviewers, who consulted a third reviewer as needed. One reviewer performed data charting with a sample of data. This was checked by a second reviewer and via discussions with the review team, as necessary.

Results: A total of 33 studies were included in this review—13 (39%) pediatric studies, 18 (55%) neonatal studies, and 2 (6%) maternal studies. All studies were published after 2011, and 27 (82%) were published from 2017 onward. The most common outcome investigated in pediatric studies was the accuracy of sepsis identification (9/13, 69%). Pediatric CCDS systems used different combinations of 18 diverse clinical criteria to detect sepsis across the 13 identified studies. In neonatal studies, 78% (14/18) of the studies investigated the Kaiser Permanente early-onset sepsis risk calculator. All studies investigated sepsis treatment and management outcomes, with 83% (15/18) reporting on antibiotics-related outcomes. Usability and cost-related outcomes were each reported in only 2 (6%) of the 31 pediatric or neonatal studies. Both studies on maternal populations were short abstracts.

Conclusions: This review found limited research investigating CCDS systems to support the early detection of sepsis among pediatric, neonatal, and maternal patients, despite the high burden of sepsis in these vulnerable populations. We have highlighted the need for a consensus definition for pediatric and neonatal sepsis and the study of usability and cost-related outcomes as critical areas for future research.

International Registered Report Identifier (IRRID): RR2-10.2196/24899

(JMIR Med Inform 2022;10(5):e35061) doi:10.2196/35061

KEYWORDS

sepsis; early detection of disease; computerized clinical decision support; patient safety; electronic health records; sepsis care pathway

Introduction

Sepsis Identification

Sepsis, redefined in adults in 2016 as "life-threatening organ dysfunction caused by a dysregulated host response to infection" [1], was associated with an estimated 11 million deaths worldwide in 2017 [2]. Neonatal, pediatric, and obstetric populations are particularly vulnerable to developing sepsis [2-4].

Children aged <5 years accounted for approximately 40% of the estimated 50 million people diagnosed with sepsis in 2017 [2]. Furthermore, a recent report indicated that children aged <1 year have a considerably higher sepsis incidence rate compared with other age groups in Australia [5]. An estimated 28 neonatal sepsis cases occur per 1000 live births, with an associated mortality rate of 17.6% [4]. Survivors of pediatric sepsis have a substantial reduction in health-related quality of life compared with nonsepsis cases, with increased risk of hospital readmissions, cognitive impairment, and physical disability [6-9]. Similarly, surviving neonatal sepsis is associated with both short- and long-term neurodevelopmental delay and disability [10,11].

The most recent consensus definition of pediatric sepsis was presented in 2005, applicable to children from full-term birth to 18 years of age, and defined pediatric sepsis as modified "systemic inflammatory response syndrome (SIRS) in the presence of or as a result of suspected or proven infection" [12]. The definition of pediatric septic shock, a severe and often fatal progression of sepsis, was refined by the 2020 Surviving Sepsis Campaign guidelines to "severe infection leading to cardiovascular dysfunction (including hypotension, need for treatment with vasoactive medication, or impaired perfusion)" [13]. There is currently no formal definition of sepsis distinct to the neonatal population [14,15]; however, a recent systematic review of randomized controlled trials found neonatal sepsis to be most commonly defined by blood culture alone, followed closely by blood culture combined with clinical signs [16].

In the maternal population, a consensus definition for maternal sepsis was presented in 2017, defined as "organ dysfunction resulting from infection during pregnancy, child-birth, post-abortion, or post-partum period" [3]. The World Health Organization Global Maternal Sepsis Study [17] found the ratio of maternal infections in hospitalized women to be 70.4 (95% CI 67.7-73.1) women per 1000 live births. Furthermore, in 2014, a World Health Organization analysis indicated that 10.7% of maternal deaths between 2003 and 2009 were associated with sepsis [18]. Maternal sepsis also affects the health of the child and has been associated with serious complications, such as neonatal sepsis, spontaneous abortions, preterm births, and over 4.5 times the risk of death in the child [3,19,20].

Prompt initiation of treatment is critical for successful sepsis management [21-23]. The earlier sepsis is detected, the faster therapies can be initiated [24]. Therefore, early detection is key to improving patient outcomes. However, pediatric, neonatal, and maternal sepsis can be challenging to identify. Age-dependent physiological norms contribute to vague or nonspecific symptoms and extreme variation between patient presentations, making it difficult for clinicians to distinguish between benign conditions and more severe disease [3,15,25-28]. Recently, clinical tools, often as part of associated care bundles and clinical programs, have been developed to facilitate improved sepsis recognition, organ dysfunction assessment, and prediction of poor outcomes for pediatric (eg, pediatric sequential organ failure assessment [29], pediatric logistic organ dysfunction-2 score [30], and pediatric sepsis score [31]), neonatal (eg, neonatal sequential organ failure assessment [32]), and maternal sepsis (eg, modified obstetric early warning score [33] and sepsis in obstetrics score [34]). However, these tools typically rely on timely and regular vital sign monitoring by clinical staff to ensure that deteriorating patients are promptly detected [35,36].

CCDS Systems

The widespread implementation of clinical information systems has allowed for sepsis recognition tools to be integrated into computerized clinician decision support (CCDS) systems [37,38] to assist clinical staff with decision-making [39]. In particular, CCDS systems can be used to improve the early detection of sepsis by monitoring patient data and automatically alerting when a patient shows signs consistent with sepsis [36]. Over the last 20 years, 2 types of CCDS systems have been developed: knowledge-based CCDS using preprogrammed rules [39] and adaptive systems using machine learning and artificial intelligence techniques [40]. This review is focused only on knowledge-based CCDS systems.

Research Questions and Aims

Despite the critical importance of sepsis detection, there is a paucity of research on pediatric, neonatal, and maternal sepsis recognition tools [14,15,17,37]. In this scoping review, we mapped the available research investigating the use of knowledge-based CCDS systems for the early detection of sepsis in pediatric, neonatal, and maternal inpatients to provide an overview of the field and identify knowledge gaps for future research. Specifically, we aimed to (1) scope the study contexts, designs, and research methods used; (2) summarize the study outcomes investigated; and (3) map the range of CCDS system designs and implementation features, such as the clinical criteria for sepsis.

Methods

Overview

A protocol detailing the methodology of this scoping review has been previously published [41]. This review follows the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews)

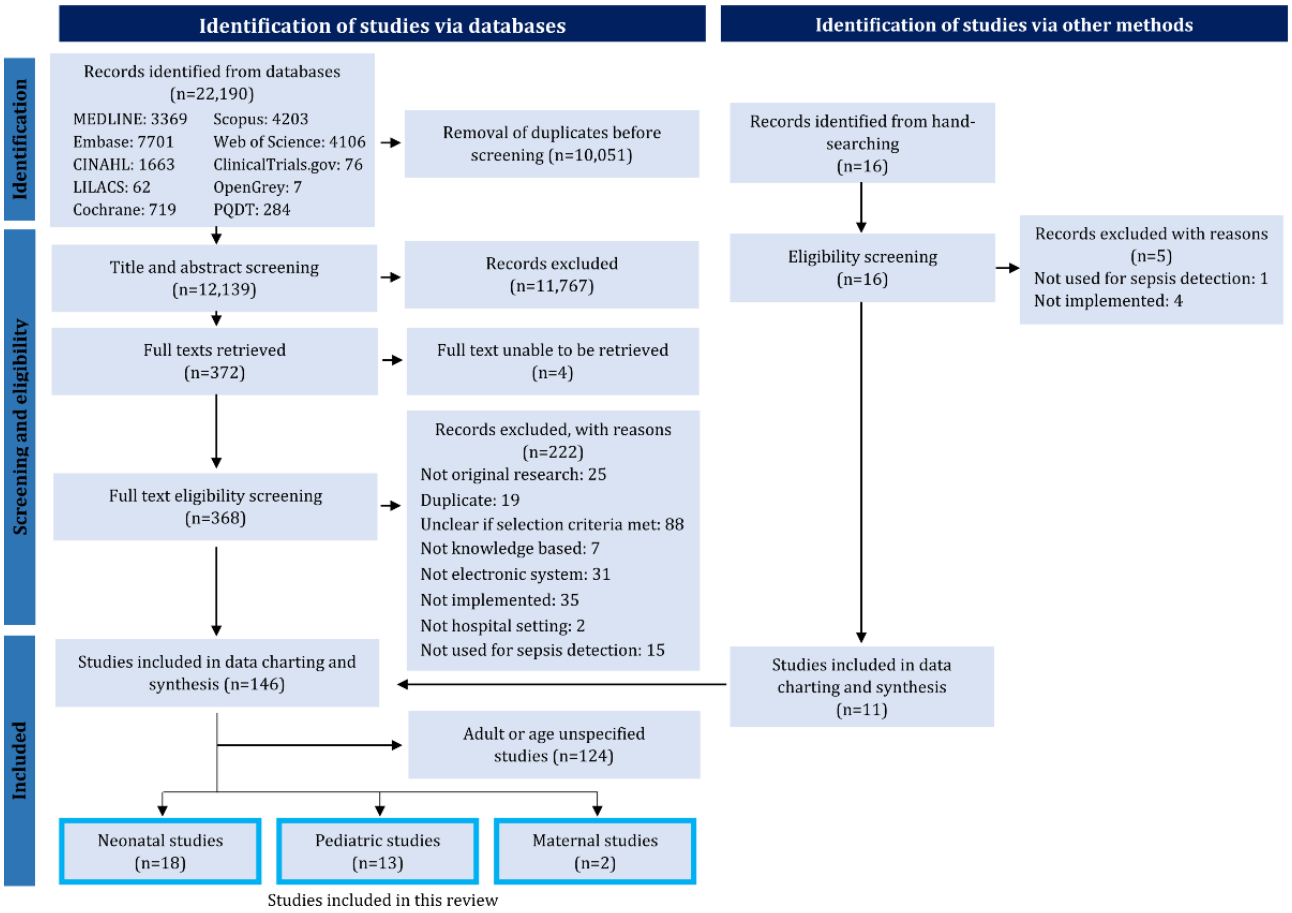
statement [42]. A completed PRISMA-ScR checklist can be found in Multimedia Appendix 1.

Study Selection

To identify relevant studies, we used a broad 3-step strategy [41], during which an experienced librarian was consulted. The final search strategy combined terms for sepsis, clinical decision support, and early detection, excluding terms for artificial intelligence, and was used to search MEDLINE, Embase, CINAHL, Cochrane, Latin American and Caribbean Health Sciences Literature (LILACS), Scopus, Web of Science, OpenGrey, ClinicalTrials.gov, and ProQuest Dissertations and Theses Global (PQDT). The search strategy used for MEDLINE is presented in Multimedia Appendix 2. The search was conducted in September 2020.

The search results were exported to an EndNote X9 (Clarivate) library. After deduplication, 2 reviewers (KA and JB) independently performed title, abstract, and full-text screening using the eligibility criteria reported in our protocol [41]. The reference lists of relevant systematic reviews and salient papers were manually searched by one reviewer (KA) with a second reviewer (JB) double-checking their inclusion to identify any further studies. Any disagreements were resolved through discussion or consultation with a third reviewer (LL). A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram visually representing this process is presented in Figure 1.

Figure 1. Flowchart of the search results and screening process. LILACS: Latin American and Caribbean Health Sciences Literature; PQDT: ProQuest Dissertations and Theses.



A total of 2 reviewers (KA and JB) independently piloted title and abstract screening with a random selection of 25 articles and full-text screening with a random selection of 10 articles. The results were discussed with a third reviewer (LL) to ensure consensus before undertaking the full screen. The 2 reviewers (KA and JB) had 100% agreement in the title and abstract pilot screen, 97.6% agreement in title and abstract screening, 60% agreement in the pilot full-text screen, and 77.4% agreement in full-text screening. Both peer-reviewed journal articles and gray literature studies, such as conference abstracts and theses, were included in this review. The gray literature that was later published as a peer-reviewed article was removed. Studies reporting the same methods and study cohorts but measuring different outcomes were included.

We chose to publish the results of this review in 2 manuscripts separated by patients' age, given the distinct sepsis presentations and pathophysiology of pediatric, neonatal, and maternal patients compared with adults [3,28,43]. The results of the review investigating adult CCDS systems have been published previously [44].

Data Charting

The form used for data charting was designed using Microsoft Access based on the data charting form previously used for adult studies [44]. The original version was refined based on

sample data extracted from 2 pediatric, 2 neonatal, and 1 maternal study. The remaining studies were charted by a single reviewer (KA), with a sample of studies checked by a second reviewer (JB), and ongoing consultation with a third reviewer (LL). We accepted any definition of the charted items, as detailed in the studies.

The final form abstracted data based on all 3 aims and included all components, as listed in our protocol [41], with some minor adjustments, as presented in Multimedia Appendix 3 [45-51]. The outcomes listed comprised (1) outcomes reported in the aims, methods, and results and (2) outcomes from the study sections that met our inclusion criteria [41]. The following were excluded: (1) outcomes mentioned in the methods or introduction but not in the results; (2) analysis of demographic or clinical features not specifically identifying the performance of the alert, unless they were the only outcome or the main outcome reported; (3) outcomes not discussed in the aims or methods and not included in the main results tables; and (4) balancing and process outcome measures. We distinguished *live* CCDS as systems that were implemented and actively alerting and *silent* CCDS as systems that were implemented and running with alerts muted.

Analyzing and Reporting the Results

The abstracted data were analyzed through a narrative review, with accompanying statistical summaries organized by population group and aims. Tables were created using frequency counts and percentages to summarize the data and produce graphical figures where appropriate. The results are presented separately for the journal articles and conference abstracts.

The charted data demonstrated substantial diversity; hence, individual categories were grouped to allow for meaningful analysis. We have included a breakdown of what is included in each group in Multimedia Appendix 4.

Ethics Approval

This scoping review used data collected from published studies (including publicly available gray literature). No individual patient was involved, and only aggregate-level data were presented; hence, ethical approval or consent to participate was not required.

Results

Study Characteristics

A database search returned 22,190 results. After deduplication, 12,139 studies were included for title and abstract screening. The full texts of 368 articles were screened, and 146 studies were identified for inclusion in the review. Manual searching identified a further 11 records. Of the 157 included studies, 33 (21%) [52-84] investigated pediatric, neonatal, and maternal populations. In comparison, 124 (79%) studies examined adult or unspecified age (assumed adult) inpatient populations (Figure 1). Thus, pediatric, neonatal, and maternal studies only represented 8.3% (13/157), 11.5% (18/157), and 1.3% (2/157) of the total studies, respectively. This process is visually presented in a PRISMA flowchart, as shown in Figure 1. A table detailing the main characteristics of the 33 included studies is presented in Multimedia Appendix 5 [52-84].

Pediatric Studies

Of the 13 studies investigating pediatric CCDS systems, 7 (54%) were journal articles and 6 (46%) were conference abstracts (Table 1). All studies were published in 2012 or later, with most journal articles (6/7, 86%) published after 2016 (Figure 2). Of the 13 studies, 11 (85%) were conducted in the United States, whereas the remaining 2 (15%) studies did not specify in which country they were conducted [64,73] (Multimedia Appendix 6). Of the 13 studies, 12 (92%) were conducted in children's hospitals, whereas the remaining study [58] was conducted at a general hospital. All studies used quantitative methods, with the principal study design split between single cohort and before-after studies (Table 1).

The most common outcomes investigated were patient outcomes and sepsis treatment and management outcomes (Figure 3). Only 1 (8%) conference abstract [58] investigated an outcome related to the CCDS system usability, and none of the studies investigated pediatric CCDS-related cost outcomes (Figure 3). The most commonly investigated patient outcome was sepsis identification (9/13, 69%; Table 1). Pediatric CCDS systems were compared with the gold standard to measure the extent to which they identified sepsis. The gold standard definition used to determine true sepsis cases differed between studies, with 13 different definitions used to define sepsis across 9 studies (Table 1). Similarly, the method used to identify gold standard cases varied across studies: 38% (5/13) performed a chart review, 8% (1/13) prospectively screened patients, 8% (1/13) applied a manual screening tool, 8% (1/13) performed both a chart review and screened patients, and 8% (1/13) did not specify.

The main characteristics of the investigated pediatric CCDS systems are presented in Table 2. Most commonly, pediatric CCDS systems were live (10/13, 77%), homegrown (11/13, 85%), alerted via the electronic health record (6/13, 46%), and responded to by nurses (6/13, 46%) and other clinicians (5/13, 38%; Table 2).

The criteria used by the CCDS systems to identify sepsis cases are summarized in Table 3. In general, a diverse range of criteria was used to identify suspected sepsis cases, with 18 clinical criteria used across 9 pediatric CCDS systems in 8 studies included in this review. The remaining 5 pediatric studies [73,74,80,82,83], all conference abstracts, did not specify the CCDS system criteria used for sepsis case identification and were not included in Table 3. A total of 2 particular systems appear to be the subject of more than one study: the first in the studies by Dewan et al [61] and Vidrine et al [81] and the second in the studies by Stinson et al [77] and Viteri et al [82]. One journal article [64] is counted twice in Table 3, as it contains 2 separate electronic CCDS systems with different criteria: one with automated continuous screening and the other with clinician-initiated screening.

Table 1. Context and outcome characteristics for pediatric studies.

Study characteristics	Number of studies by publication		Total[a]
	Journal articles	Conference abstracts	
Subtotal, n	7	6	13
Principal study type, n (%)			
Single cohort	3 (43)	4 (67)	7 (54)
Before-after	4 (57)	2 (33)	6 (46)
Setting, n (%)			
Hospital wide[b]	0 (0)	2 (33)	2 (15)
Emergency department	4 (57)	1 (17)	5 (38)
Intensive care unit	2 (29)	0 (0)	2 (15)
Inpatient units	1 (14)	3 (50)	4 (31)
Number of participants, n (%)			
≤100	1 (14)	2 (33)	3 (23)
101-10,000	1 (14)	2 (33)	3 (23)
10,001-100,000	2 (29)	1 (17)	3 (23)
>100,000	2 (29)	0 (0)	2 (15)
Unspecified	1 (14)	1 (17)	2 (15)
Funding, n (%)			
Yes (noncommercial)	2 (29)	0 (0)	2 (15)
No	2 (29)	0 (0)	2 (15)
Unspecified	3 (43)	6 (100)	9 (69)
Outcomes, n (%)			
Patient outcomes			
Sepsis identification	5 (71)	4 (67)	9 (69)
Gold standard definition[c]			
Goldstein et al [12]	2 (29)	0 (0)	2 (15)
American Academy of Pediatrics Sepsis Collaborative tool [85]	1 (14)	0 (0)	1 (8)
Clinician discretion	3 (43)	2 (33)	5 (38)
Improving Pediatric Sepsis Outcomes definition [86]	1 (14)	0 (0)	1 (8)
International Classification of Diseases codes	1 (14)	0 (0)	1 (8)
Not specified	1 (14)	2 (33)	3 (23)
Other	4 (57)	1 (17)	5 (38)
Sepsis treatment or management, n (%)			
Timeliness of alert or intervention	3 (43)	1 (17)	4 (31)
Other	6 (86)	1 (17)	7 (54)
Usability, n (%)			
Satisfaction	0 (0)	1 (17)	1 (8)

[a]The percentages were calculated from the number of pediatric studies (n=13). As some studies reported multiple outcomes for each category, there were more than 13 outcomes in some categories, and therefore, the percentages add to more than 100%.

[b]If the study setting was not explicitly stated, it was assumed to be hospital wide.

[c]Some studies have used multiple definitions of sepsis as part of their gold standard.

Figure 2. Studies investigating neonatal and pediatric computerized clinician decision support systems by year, population, and publication type.

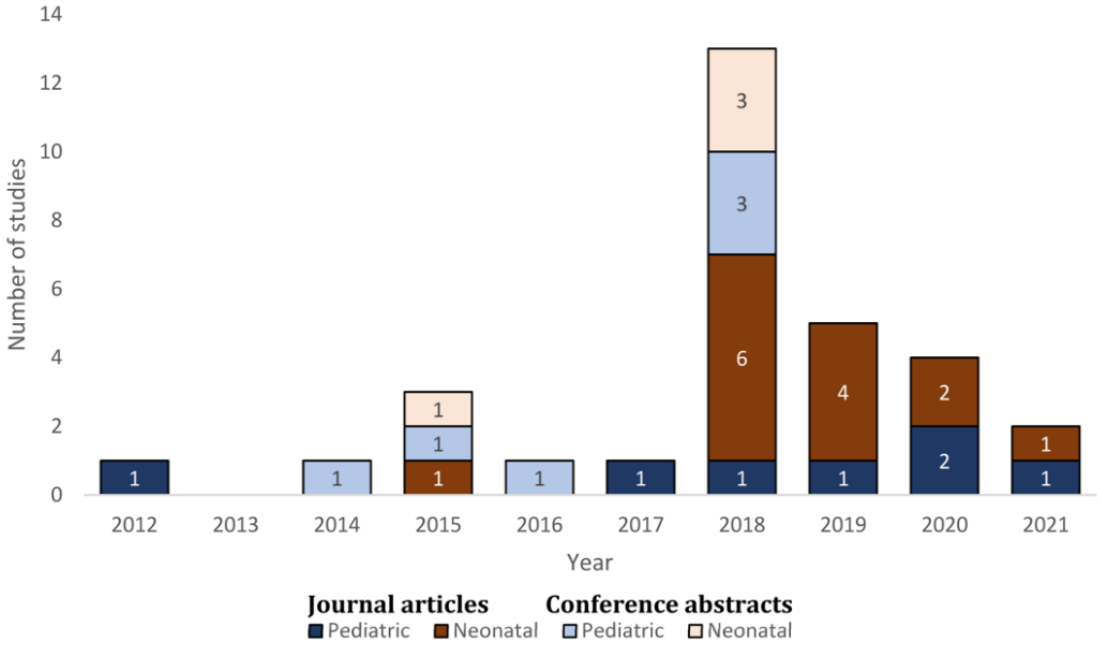


Figure 3. Outcome categories reported by studies by publication type and population.

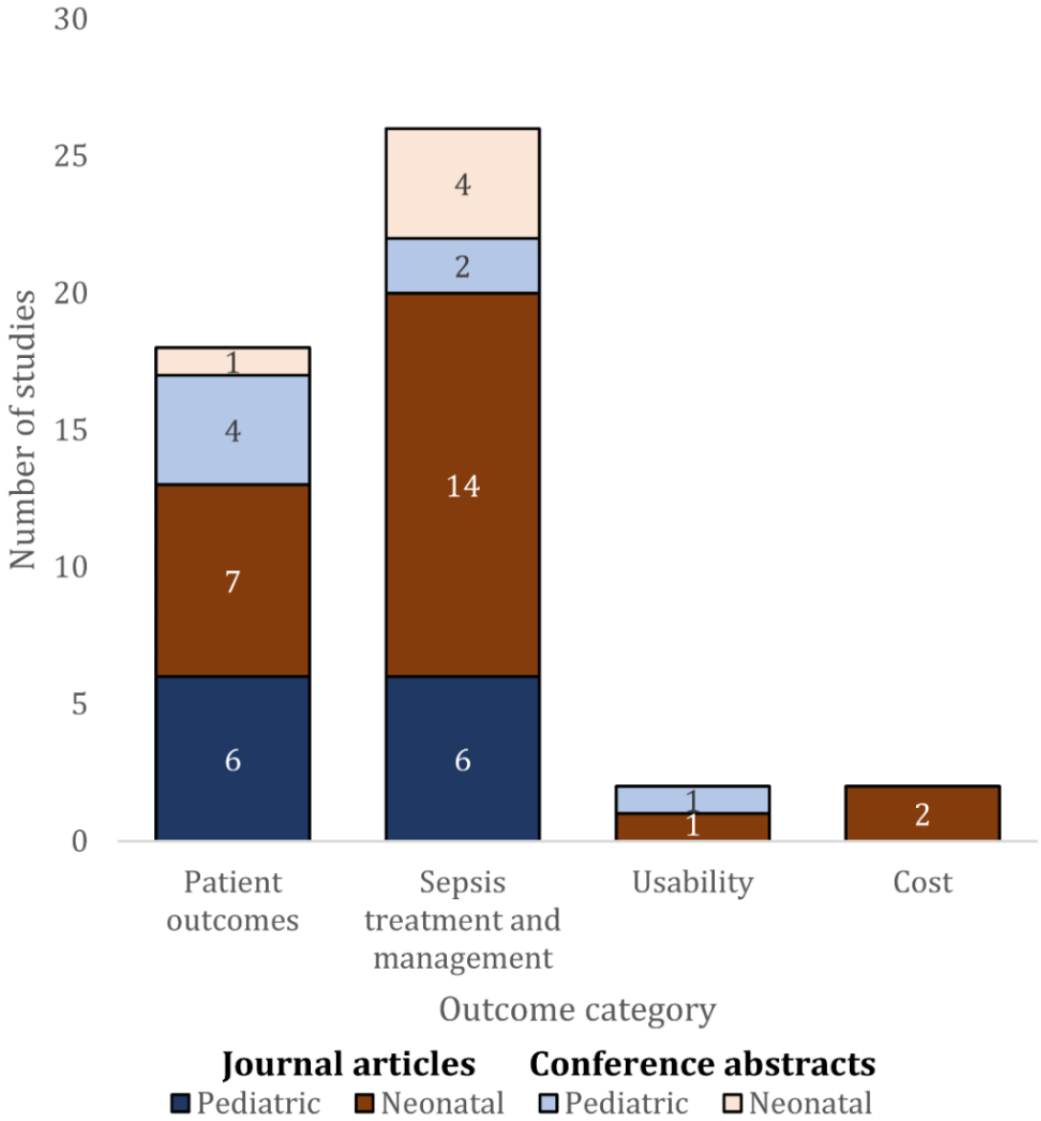


Table 2. Computerized clinical decision support characteristics in pediatric studies.

CCDS[a] characteristics	Number of studies by publication		Total[b]
	Journal articles	Conference abstracts	
Subtotal, n	7	6	13
CCDS type, n (%)			
Homegrown[c]	6 (86)	5 (83)	11 (85)
Commercial, n (%)	0 (0)	1 (17)	1 (8)
Epic monitor	0 (0)	1 (17)	1 (8)
Unspecified	1 (14)	0 (0)	1 (8)
Silent or live[d], n (%)			
Live	5 (71)	5 (83)	10 (77)
Silent	1 (14)	1 (17)	2 (15)
Both (pre or post)	1 (14)	0 (0)	1 (8)
Related interventions, n (%)			
None	2 (29)	5 (83)	7 (54)
Response team	4 (57)	1 (17)	5 (38)
Education and information resources	3 (43)	1 (17)	4 (31)
Order sets	3 (43)	1 (17)	4 (31)
Sepsis protocol	1 (14)	1 (17)	2 (15)
Other	2 (29)	1 (17)	3 (23)
Responding personnel, n (%)			
Nurses	6 (86)	0 (0)	6 (46)
Other clinicians	3 (43)	2 (33)	5 (38)
Response team	0 (0)	2 (33)	2 (15)
Not specified	0 (0)	3 (50)	3 (23)
Alert delivery, n (%)			
Electronic health record	6 (86)	0 (0)	6 (46)
Emergency department tracking board	1 (14)	0 (0)	1 (8)
Not specified	0 (0)	6 (100)	6 (46)

[a]CCDS: computerized clinical decision support.

[b]The percentages were calculated from the number of pediatric studies (n=13). As some studies reported multiple characteristics for each category, there were more than 13 characteristics in some categories; therefore, the percentages add to more than 100%.

[c]Homegrown CCDS systems are defined as CCDS systems that have been designed by the institution implementing them, rather than commercially available systems [41].

[d]A *live* CCDS system is a system that is implemented and being used by clinicians in real time during the study. Silent systems are systems that have been implemented but do not alert clinicians during the study and thus do not influence treatment.

Table 3. Clinical criteria used by pediatric computerized clinical decision support (CCDS) systems for sepsis identification.

	Study									
	Balamuth et al, 2017 [55]	Cruz et al, 2012 [59]	Dewan et al, 2020 [61]	Eisenberg et al, 2021 [64] (clinician-initiated)	Eisenberg et al, 2021 [64] (automated)	Lloyd et al, 2018 [71]	Stinson et al, 2019 [77]	Vidrine et al, 2020 [81]	Coffman et al, 2018[a] [58]	Total, n (%[b])
Temperature	✓	✓	✓	✓	✓	✓	✓	✓	✓	9 (69)
Capillary refill or perfusion	✓	✓	✓	✓		✓	✓	✓		7 (54)
Mental status	✓	✓	✓	✓		✓	✓	✓		7 (54)
Heart rate	✓	✓		✓	✓	✓	✓			6 (46)
Hypotension	✓		✓	✓		✓	✓	✓		6 (46)
High-risk patient	✓	✓		✓		✓	✓			5 (38)
Pulse assessment			✓	✓		✓	✓	✓		5 (38)
Skin assessment			✓	✓		✓	✓	✓		5 (38)
Respiratory rate				✓	✓	✓	✓			4 (31)
Infection concern, change in clinical or sepsis risk	✓			✓		✓				3 (23)
Blood culture order			✓					✓		2 (15)
Leukocyte count					✓					1 (8)
Cardiac organ dysfunction					✓					1 (8)
Noncardiac organ dysfunction					✓					1 (8)
Change in Pediatric Early Warning Score									✓	1 (8)
Family concern									✓	1 (8)
Vital sign change									✓	1 (8)
Patient risk change									✓	1 (8)

[a]This study is a conference abstract, and the other 8 studies are journal articles.

[b]The percentages were calculated from the number of pediatric studies (n=13).

Neonatal Studies

Of the 18 articles investigating neonatal CCDS systems, 14 (78%) were journal articles and 4 (22%) were conference abstracts. All studies were published in 2015 or later, with most published in 2018 (n=9; Figure 2). Overall, 61% (11/18) of the studies were conducted in the United States, 11% (2/18) were conducted in the Netherlands, 11% (2/18) did not specify location, and 1 (6%) study each was set in Australia, Israel, and the United Kingdom (Multimedia Appendix 6). All neonatal studies used quantitative methods to investigate the CCDS systems. A total of 89% (16/18) of studies were single site, with the remaining 11% (2/18) of studies involving 4 [66] and 2 sites [69]. The gestational age range of neonates included in these studies was quite diverse, with 35 weeks and older being the most common inclusion threshold (Table 4).

The most common outcome used to investigate neonatal CCDS systems was sepsis treatment and management outcomes, followed by patient outcomes (Figure 3; Table 4). CCDS-related usability and cost outcomes were only investigated by 1 and 2 studies, respectively [53,56,66] (Figure 3; Table 4). Of the sepsis treatment and management outcomes, antibiotics-related outcomes were reported most frequently (15/18, 83%; Table 4). Table 5 reports the main characteristics of the neonatal CCDS systems. Notably, most studies investigated early-onset sepsis (15/18, 83%) using the neonatal early-onset sepsis risk calculator developed by the Kaiser Permanente team [87-89] (14/18, 78%).

Table 4. Context and outcome characteristics in neonatal studies.

Study characteristics	Number of studies by publication		Total[a]
	Journal articles	Conference abstracts	
Subtotal, n	14	4	18
Principal study type, n (%)			
Single cohort	3 (21)	3 (75)	6 (33)
Before-after	9 (64)	1 (25)	10 (56)
Interrupted time series	2 (14)	0 (0)	2 (11)
Setting, n (%)			
Hospital wide[b]	4 (29)	2 (50)	6 (33)
Nursery	7 (50)	2 (50)	9 (50)
ICU[c]	3 (21)	0 (0)	3 (17)
Number of participants, n (%)			
≤100	0 (0)	1 (25)	1 (6)
101-1000	5 (36)	1 (25)	6 (33)
1001-10,000	6 (43)	0 (0)	6 (33)
>10,001	2 (14)	0 (0)	2 (11)
Unspecified	1 (7)	2 (50)	3 (17)
Age of included neonates, n (%)			
<33 weeks gestation	1 (7)	0 (0)	1 (6)
≥34 weeks gestation	3 (21)	1 (25)	4 (22)
≥35 weeks gestation	4 (29)	1 (25)	5 (28)
≥36 weeks gestation	2 (14)	0 (0)	2 (11)
>37 weeks gestation	1 (7)	0 (0)	1 (6)
First month of life	1 (7)	0 (0)	1 (6)
Unspecified	2 (14)	2 (50)	4 (22)
Funding, n (%)			
Yes (noncommercial)	1 (7)	0 (0)	1 (6)
No	7 (50)	0 (0)	7 (39)
Unspecified	6 (43)	4 (100)	10 (56)
Outcomes, n (%)			
Patient outcomes			
ICU admission	4 (29)	0 (0)	4 (22)
Length of stay	3 (21)	1 (25)	4 (22)
Other	4 (29)	1 (25)	5 (28)
Sepsis treatment or management			
Antibiotics	12 (86)	3 (75)	15 (83)
Laboratory evaluation	8 (57)	3 (75)	11 (61)
Timeliness of alert or intervention	2 (14)	0 (0)	2 (11)
Sepsis guideline compliance	2 (14)	0 (0)	2 (11)
Other	4 (29)	1 (25)	5 (28)
Usability			
Effectiveness	1 (7)	0 (0)	1 (6)

Study characteristics	Number of studies by publication		Total[a]
	Journal articles	Conference abstracts	
Cost	2 (14)	0 (0)	2 (11)

[a]The percentages were calculated from the number of neonatal studies (n=18). As some studies have reported multiple outcomes for each category, there were more than 18 outcomes in some categories; therefore, the percentages add to more than 100%.

[b]If the study setting was not explicitly stated, it was assumed to be hospital wide.

[c]ICU: intensive care unit.

Table 5. Computerized clinical decision support characteristics in neonatal studies.

CCDS[a] characteristics	Number of studies by publication		Total[b]
	Journal articles	Conference abstracts	
Subtotal, n	14	4	18
Type of sepsis, n (%)			
Early-onset sepsis	12 (86)	3 (75)	15 (83)
Late-onset sepsis	1 (7)	0 (0)	1 (6)
Sepsis	1 (7)	1 (25)	2 (11)
General CCDS criteria, n (%)			
Kaiser Permanente early-onset sepsis risk [89]	12 (86)	2 (50)	14 (78)
Epic Monitor [65]	1 (7)	1 (25)	2 (11)
RALIS [69]	1 (7)	0 (0)	1 (6)
Not specified	0 (0)	1 (25)	1 (6)
Silent or live[c], n (%)			
Live	12 (86)	4 (100)	16 (89)
Silent	1 (7)	0 (0)	1 (6)
Both (pre or post)	1 (7)	0 (0)	1 (6)
Related interventions, n (%)			
Education and information resources	8 (57)	1 (25)	9 (50)
None	4 (29)	3 (75)	7 (39)
Sepsis protocol	4 (29)	0 (0)	4 (22)
Order sets	2 (14)	0 (0)	2 (11)
Other	5 (36)	1 (25)	6 (33)
Responding personnel, n (%)			
Nurses	4 (29)	1 (25)	5 (28)
Other clinicians	10 (71)	0 (0)	10 (56)
Paramedics	1 (7)	1 (25)	2 (11)
Not specified	2 (14)	2 (50)	4 (22)
Alert delivery, n (%)			
Calculated by personnel	10 (71)	3 (75)	13 (72)
Other	2 (14)	0 (0)	2 (11)
Not specified	2 (14)	1 (25)	3 (17)

[a]CCDS: computerized clinical decision support.

[b]The percentages were calculated from the number of neonatal studies (n=18). As some studies have reported multiple characteristics for each category, there were more than 18 characteristics, therefore, the percentages add to more than 100%.

[c]A *live* CCDS system is a system that is implemented and being used by clinicians in real time during the study. Silent systems are systems that have been implemented but do not alert clinicians during the study and thus do not influence treatment.

Maternal Studies

Only 2 studies—those by Davis et al [60] and Blumenthal et al [57]—have investigated CCDS systems for sepsis in pregnant or immediately postpartum populations. Both studies were abstracts and used quantitative methods. Blumenthal et al [57] used a before-after study design, whereas Davis et al [60] did not provide sufficient information for the study design to be determined. Davis et al [60] conducted a single-site, hospital-wide study in the United States, and Blumenthal et al [57] conducted a study at 3 sites but did not specify in which country. None of the studies reported on the number of participants. To identify maternal sepsis, Davis et al [60] used the obstetric-adjusted systemic inflammatory response syndrome (SIRS) criteria (comprising SIRS with the addition of fetal heart rate) plus organ dysfunction, whereas Blumenthal et al [57] used a maternal early warning score (comprising temperature plus heart rate, altered mental state, respiratory rate, and mean arterial pressure). Both studies investigated sepsis treatment and management outcomes, with Blumenthal et al [57] additionally investigating patient outcomes.

Discussion

Principal Findings

This review comprehensively scoped the current literature on CCDS systems for early detection of sepsis in pediatric, neonatal, and maternal hospital populations. Overall, our findings highlight the scarcity of studies in these unique populations when compared with the general adult population, representing only 21% (33/157) of studies. Furthermore, only 64% (21/33) of studies were peer-reviewed journal articles. Given the high burden of sepsis in pediatric, neonatal, and maternal patients, this comparatively small number of studies is concerning [2-4,18] and underlines the critical need for future high-quality research into CCDS systems for these vulnerable populations. However, the rapid expansion of this field in recent years is encouraging, with all 33 studies published in the last 10 years and the majority (26/33, 79%) published in the last 5 years.

Pediatric Sepsis

Our findings emphasize the variability in pediatric studies that have evaluated the use of sepsis CCDS systems. In particular, we found great variability across the clinical criteria used for pediatric sepsis identification, with 18 different clinical criteria used in numerous combinations across 8 studies (Table 3). Furthermore, a range of gold standard definitions was applied, of which the most common was clinician discretion rather than published tools [12,85,86], highlighting the lack of a consensus definition and tool for pediatric sepsis identification. Hospital settings varied widely between studies, and numerous related interventions were implemented alongside the pediatric CCDS, with few similarities. This variability makes it difficult to compare studies and draw generalized conclusions from the literature. All studies were single cohort or before-after studies, highlighting the need for more robust study designs to provide stronger evidence regarding the use of CCDS systems.

The heterogeneity in the clinical criteria used, both for the CCDS system and the gold standard definitions, can be attributed to a lack of current consensus regarding pediatric identification, risk stratification, and diagnosis. Although the definition of adult sepsis was updated in 2016 [1], followed by the publication of the quick sepsis-related organ failure assessment tool [90], the most recent pediatric sepsis consensus definition was in 2005 [12] and has exhibited numerous limitations [31,91,92]. An extensive study by Weiss et al [93] found an interrater agreement of only 0.57 between the 2005 consensus and physician diagnosis of pediatric sepsis, further emphasizing the inadequacies of the current consensus criteria in practice. Researchers have since attempted to adapt the quick sepsis-related organ failure assessment to the pediatric population or pediatric logistic organ dysfunction-2, a pediatric deterioration tool, to sepsis [29,30,94,95]. Preliminary results from these studies show promise, demonstrating moderate to high prognostic accuracy for poor patient outcomes, such as mortality and pediatric intensive care unit admission [29,30,94,95]. Critical to this challenge is the unique pathophysiology of pediatric sepsis, in which simply age-adjusting adult sepsis criteria is controversial and inadequate [91,96]. For example, hypotension is commonly used as a key indicator of septic shock in adults; however, it is less useful in children, as hypotension is typically not present until much later in the disease course [25,26,91]. In addition, symptoms considered key to adult sepsis identification, such as tachycardia and tachypnea, are common in febrile children regardless of disease severity and can often be present due to crying and distress [25,26,95]. Therefore, there have been numerous calls by both academics and clinicians for an updated pediatric consensus in recent years [13,43,91,95]. In 2019, the Society of Critical Care Medicine convened the Pediatric Sepsis Definition Taskforce to update the consensus criteria for pediatric sepsis identification [97]. Although they have recently published a systematic review investigating the individual factors, clinical criteria, or illness severity scores that are used to identify children with sepsis who are at higher risk of developing organ dysfunction or death, the task force has not yet released an updated definition [97]. The absence of an up-to-date consensus for defining or detecting pediatric sepsis has likely contributed to the high diversity of CCDS clinical criteria used in pediatric populations and the range of definitions used for gold standard pediatric sepsis detection. Our findings demonstrate the need for more robust evidence to investigate the appropriate clinical criteria for pediatric sepsis and reinforce the urgent need for an updated consensus on the definition of pediatric sepsis.

Notably, an updated pediatric consensus must consider the extensive chronological and developmental age-dependent variability found in the pediatric population. For example, the pathophysiology of sepsis is expected to differ significantly among an adolescent, a child aged 5 years, and an infant aged 2 months. This will likely affect how different pediatric age groups present with sepsis, and accounting for these changes may not be as simple as adjusting the normal threshold of different vital signs according to age. This diversity needs to be studied and reflected in future consensus definitions and clinical criteria of the CCDS system.

Neonatal Sepsis

Our findings report considerable variation across neonatal studies, despite most studies evaluating the same CCDS system: the Kaiser Permanente early-onset sepsis risk calculator (KPC) [89]. In particular, the gestational age of the neonates included in the study varied considerably (Table 4). Most studies investigated moderate to late preterm and term infants, with cutoffs for gestational age ranging from ≥34 to >37 weeks [98] or infants within their first month of life. A single study [69] investigated very preterm infants at <33 weeks gestational age [98], indicating a key research gap, as preterm infants are at a considerably higher risk of sepsis and infection than full-term newborns [14,28,32,99]. A recent study [99] demonstrated that more than one-third (38%) of extremely preterm infants, defined as infants ≤28 weeks' gestation, had late-onset sepsis. The included studies investigated a diverse range of outcomes, related interventions, and responding personnel. Large multisite studies would improve the generalizability of the literature and thus should be considered despite the substantial difficulty in undertaking them.

Of the 18 neonatal studies included in this review, 14 (78%) investigated KPC [89]. This calculator combines the baseline early-onset sepsis incidence with maternal and infant characteristics and a clinical evaluation [89]. It aims to identify neonates at risk of early-onset sepsis, defined as sepsis within the first 72 hours after birth [28,87,88]. Under conventional sepsis management guidelines, many neonates are given potentially unnecessary antibiotic therapy as a precaution against sepsis, resulting in unintended negative effects [14,87]. A systematic review and meta-analysis performed by Achten et al [100] demonstrated that the use of KPC was associated with a reduction in antibiotic use. However, a more recent meta-analysis [101] showed that the KPC missed many cases of early-onset sepsis compared with the UK National Institute for Health and Care Excellence guidelines. This results in delayed or missed treatment for these neonates and suggests that further evaluation of the calculator is required [101]. In addition, the KPC is only designed for predicting sepsis risk in infants born at ≥34 weeks' gestation within a very narrow early-onset sepsis time frame [87-89]. Our review identified only 17% (3/18) of neonatal studies that did not examine early-onset sepsis, with 6% (1/18) investigating late-onset sepsis and 11% (2/18) investigating general neonatal sepsis. Late-onset neonatal sepsis, often defined as sepsis occurring ≥3 days after birth, is a leading cause of mortality in vulnerable preterm infants [28,32,99,102]. This calls attention to a clear knowledge gap for future research into CCDS systems for neonatal sepsis occurring outside the initial 72 hours of life.

To date, no consensus definition has been developed for neonatal sepsis [15,16,28,103]. As the neonatal population is uniquely different from adults and older children, current adult and pediatric clinical criteria cannot be simply adapted [15,32,103]. A recently published systematic review [16] highlighted the variance in the currently used definitions of neonatal sepsis in randomized controlled trials. Surprisingly, the most commonly used definition was microbiological culture by itself or in combination with clinical signs and symptoms, despite the proven low sensitivity of this method and the high incidence of culture-negative sepsis among the neonatal population [14,16,102]. Similarly, some studies included in this review required a positive culture test to diagnose neonatal sepsis. A consensus on the definition of neonatal sepsis is needed to better identify suspected neonatal sepsis in clinical practice, for research studies, and to improve antibiotic stewardship in newborns [14,15,28,103]. Furthermore, any consensus criterion must acknowledge the age-related variability inherent to the neonatal population, as sepsis pathophysiology differs considerably between a preterm neonate and an infant in their first month of life [103].

Maternal Sepsis

Despite the devastating consequences of sepsis in pregnant and immediately postpartum women [3,17,18], our comprehensive literature search identified only 2 studies that evaluated the use of CCDS systems for maternal sepsis. Pregnancy involves extensive physiological, hormonal, and psychological changes, which may mask the common symptoms of sepsis, resulting in delayed diagnosis and treatment [3,19,104]. A systematic review by Bauer et al [104] demonstrated that healthy pregnant women during the second and third trimesters often demonstrate considerable overlap with the SIRS criteria. This alteration of the usual physiological state must be represented in CCDS systems to ensure that sepsis in pregnant and immediately postpartum women is detected early, without the risk of unnecessary treatment in healthy patients. The lack of high-quality peer-reviewed studies in this population underlines a concerning knowledge gap in the literature, for which further research is urgently needed.

Usability and Cost of CCDS Systems

The usability of any health intervention technology is critical for its successful implementation [105-108]. Therefore, investigating the usability of CCDS systems is essential for developing efficient and functional systems. In particular, alarm fatigue is a well-established usability concern for CCDS systems [109]. Alarm fatigue occurs when clinicians become desensitized to frequent inappropriate alarms and begin ignoring or overriding alerts, reducing the effectiveness of alert systems and potentially impacting patient outcomes [109,110]. To prevent alarm fatigue, CCDS systems must be carefully calibrated to avoid unnecessary frequent alerting [109,110]. None of the studies reported in this review investigated alarm fatigue in response to the implemented CCDS system, despite its importance for successful CCDS use.

Understanding the cost or cost-effectiveness of an intervention supports policy and clinical decision-making when determining resource allocation under limited health care budgets [111]. This is especially true for sepsis, which represents a large financial burden on the health system through both acute hospital care and long-term treatment and rehabilitation [112,113]. Of the 33 studies included in this review, only 4 (12%) investigated outcomes related to cost or usability, 1 (3%) in pediatric and 3 (9%) in neonatal populations, demonstrating a clear evidence gap for future research.

Strengths and Limitations

This review comprehensively searched the available literature, both peer reviewed and gray, on the use of CCDS systems for inpatients with neonatal, pediatric, and maternal sepsis. Owing to time and resource constraints, the searches were limited to studies available in the English language and thus may have missed publications in other languages. Furthermore, the data extraction was performed by only 1 reviewer (KA). To limit any consequential data entry errors, the extraction form was extensively piloted, and any issues were cross-checked and fully discussed with the review team.

Conclusions

Our findings have illustrated a comparative scarcity of studies investigating CCDS systems in pediatric, neonatal, and maternal inpatients, despite their high sepsis burden. Further research is needed to evaluate CCDS systems for the early detection of sepsis in these vulnerable populations. We identified extensive variation in the clinical criteria and gold standard definitions used by pediatric CCDS systems, and our findings reinforce calls for updated pediatric and neonatal sepsis consensus definitions. The review also shows a clear absence of studies investigating CCDS systems for sepsis identification in maternal inpatients, high-risk preterm populations, and neonates outside the first 72 hours of life. Finally, our review demonstrated a lack of studies investigating the usability and cost of CCDS systems, both of which are key to their effectiveness and sustainability. In conclusion, our review has identified substantial and important knowledge gaps in the literature evaluating CCDS systems for the early detection of sepsis in pediatric, neonatal, and maternal populations, which would benefit greatly from future research.

Acknowledgments

The authors would like to thank Mr Jeremy Cullis, an experienced clinical librarian, for his expert advice and assistance in designing our final MEDLINE search strategy and in translating it for other databases. The authors have received no financial support or funding for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 516 KB - medinform_v10i5e35061_app1.pdf]

Multimedia Appendix 2
MEDLINE search strategy.
[PDF File (Adobe PDF File), 75 KB - medinform_v10i5e35061_app2.pdf]

Multimedia Appendix 3
Adjustments made to data charting form.
[PDF File (Adobe PDF File), 283 KB - medinform_v10i5e35061_app3.pdf]

Multimedia Appendix 4
Definitions of categories combining multiple subgroups.
[PDF File (Adobe PDF File), 175 KB - medinform_v10i5e35061_app4.pdf]

Multimedia Appendix 5
Main study characteristics table.
[PDF File (Adobe PDF File), 238 KB - medinform_v10i5e35061_app5.pdf]

Multimedia Appendix 6
Study setting by country.
[PDF File (Adobe PDF File), 122 KB - medinform_v10i5e35061_app6.pdf]

References

1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810 [FREE Full text] [doi: 10.1001/jama.2016.0287] [Medline: 26903338]

2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. Lancet 2020 Jan 18;395(10219):200-211 [FREE Full text] [doi: 10.1016/S0140-6736(19)32989-7] [Medline: 31954465]
3. Bonet M, Nogueira Pileggi V, Rijken MJ, Coomarasamy A, Lissauer D, Souza JP, et al. Towards a consensus definition of maternal sepsis: results of a systematic review and expert consultation. Reprod Health 2017 May 30;14(1):67 [FREE Full text] [doi: 10.1186/s12978-017-0321-6] [Medline: 28558733]
4. Fleischmann C, Reichert F, Cassini A, Horner R, Harder T, Markwart R, et al. Global incidence and mortality of neonatal sepsis: a systematic review and meta-analysis. Arch Dis Child 2021 Jan 22;106(8):745-752 [FREE Full text] [doi: 10.1136/archdischild-2020-320217] [Medline: 33483376]
5. Li L, Sunderland N, Rathnayake K, Westbrook JI. Epidemiology of Sepsis in Australian Public Hospitals: A Mixed Methods, National Longitudinal Study (2013-2018). ACSQHC. 2020. URL: https://www.safetyandquality.gov.au/sites/default/files/2020-05/epidemiology_of_sepsis_-_february_2020_002.pdf [accessed 2021-11-03]
6. Killien EY, Farris RW, Watson RS, Dervan LA, Zimmerman JJ. Health-related quality of life among survivors of pediatric sepsis. Pediatr Crit Care Med 2019 Jun;20(6):501-509 [FREE Full text] [doi: 10.1097/PCC.0000000000001886] [Medline: 30720672]
7. Weiss SL, Fitzgerald JC, Pappachan J, Wheeler D, Jaramillo-Bustamante JC, Salloo A, Sepsis Prevalence, Outcomes,Therapies (SPROUT) Study InvestigatorsPediatric Acute Lung InjurySepsis Investigators (PALISI) Network. Global epidemiology of pediatric severe sepsis: the sepsis prevalence, outcomes, and therapies study. Am J Respir Crit Care Med 2015 May 15;191(10):1147-1157 [FREE Full text] [doi: 10.1164/rccm.201412-2323OC] [Medline: 25734408]
8. Farris RW, Weiss NS, Zimmerman JJ. Functional outcomes in pediatric severe sepsis: further analysis of the researching severe sepsis and organ dysfunction in children: a global perspective trial. Pediatr Crit Care Med 2013 Nov;14(9):835-842 [FREE Full text] [doi: 10.1097/PCC.0b013e3182a551c8] [Medline: 24108117]
9. Prout AJ, Talisa VB, Carcillo JA, Angus DC, Chang CH, Yende S. Epidemiology of readmissions after sepsis hospitalization in children. Hosp Pediatr 2019 Apr;9(4):249-255 [FREE Full text] [doi: 10.1542/hpeds.2018-0175] [Medline: 30824488]
10. Savioli K, Rouse C, Susi A, Gorman G, Hisle-Gorman E. Suspected or known neonatal sepsis and neurodevelopmental delay by 5 years. J Perinatol 2018 Nov;38(11):1573-1580. [doi: 10.1038/s41372-018-0217-5] [Medline: 30202045]
11. Cai S, Thompson DK, Anderson PJ, Yang JY. Short- and long-term neurodevelopmental outcomes of very preterm infants with neonatal sepsis: a systematic review and meta-analysis. Children (Basel) 2019 Dec 01;6(12):131 [FREE Full text] [doi: 10.3390/children6120131] [Medline: 31805647]
12. Goldstein B, Giroir B, Randolph A, International Consensus Conference on Pediatric Sepsis. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. Pediatr Crit Care Med 2005 Jan;6(1):2-8. [doi: 10.1097/01.PCC.0000149131.72248.E6] [Medline: 15636651]
13. Weiss SL, Peters MJ, Alhazzani W, Agus MS, Flori HR, Inwald DP, et al. Surviving sepsis campaign international guidelines for the management of septic shock and sepsis-associated organ dysfunction in children. Pediatr Crit Care Med 2020 Feb;21(2):e52-106. [doi: 10.1097/PCC.0000000000002198] [Medline: 32032273]
14. Wynn JL, Polin RA. Progress in the management of neonatal sepsis: the importance of a consensus definition. Pediatr Res 2018 Jan;83(1-1):13-15. [doi: 10.1038/pr.2017.224] [Medline: 29019470]
15. McGovern M, Giannoni E, Kuester H, Turner MA, van den Hoogen A, Bliss JM, Infection, Inflammation, ImmunologyImmunisation (I4) section of the ESPR. Challenges in developing a consensus definition of neonatal sepsis. Pediatr Res 2020 Jul;88(1):14-26. [doi: 10.1038/s41390-020-0785-x] [Medline: 32126571]
16. Hayes R, Hartnett J, Semova G, Murray C, Murphy K, Carroll L, Infection, Inflammation, ImmunologyImmunisation (I4) section of the European Society for Paediatric Research (ESPR). Neonatal sepsis definitions from randomised clinical trials. Pediatr Res 2021 Nov 06 (forthcoming). [doi: 10.1038/s41390-021-01749-3] [Medline: 34743180]
17. WHO Global Maternal Sepsis Study (GLOSS) Research Group. Frequency and management of maternal infection in health facilities in 52 countries (GLOSS): a 1-week inception cohort study. Lancet Glob Health 2020 May;8(5):e661-e671 [FREE Full text] [doi: 10.1016/S2214-109X(20)30109-1] [Medline: 32353314]
18. Say L, Chou D, Gemmill A, Tunçalp Ö, Moller A, Daniels J, et al. Global causes of maternal death: a WHO systematic analysis. Lancet Glob Health 2014 Jun;2(6):e323-e333 [FREE Full text] [doi: 10.1016/S2214-109X(14)70227-X] [Medline: 25103301]
19. Escobar MF, Echavarría MP, Zambrano MA, Ramos I, Kusanovic JP. Maternal sepsis. Am J Obstet Gynecol MFM 2020 Aug;2(3):100149. [doi: 10.1016/j.ajogmf.2020.100149] [Medline: 33345880]
20. Scott S, Kendall L, Gomez P, Howie SR, Zaman SM, Ceesay S, et al. Effect of maternal death on child survival in rural West Africa: 25 years of prospective surveillance data in The Gambia. PLoS One 2017;12(2):e0172286 [FREE Full text] [doi: 10.1371/journal.pone.0172286] [Medline: 28225798]
21. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med 2017 Jun 08;376(23):2235-2244 [FREE Full text] [doi: 10.1056/NEJMoa1703058] [Medline: 28528569]

22. Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. Crit Care Med 2017 Mar;45(3):486-552. [doi: 10.1097/CCM.0000000000002255] [Medline: 28098591]
23. Weiss SL, Fitzgerald JC, Balamuth F, Alpern ER, Lavelle J, Chilutti M, et al. Delayed antimicrobial therapy increases mortality and organ dysfunction duration in pediatric sepsis. Crit Care Med 2014 Nov;42(11):2409-2417 [FREE Full text] [doi: 10.1097/CCM.0000000000000509] [Medline: 25148597]
24. Cecconi M, Evans L, Levy M, Rhodes A. Sepsis and septic shock. Lancet 2018 Jul 07;392(10141):75-87. [doi: 10.1016/S0140-6736(18)30696-2] [Medline: 29937192]
25. Schlapbach LJ, Weiss SL, Wolf J. Reducing collateral damage from mandates for time to antibiotics in pediatric sepsis-primum non nocere. JAMA Pediatr 2019 May 01;173(5):409-410. [doi: 10.1001/jamapediatrics.2019.0174] [Medline: 30882879]
26. Cruz AT, Lane RD, Balamuth F, Aronson PL, Ashby DW, Neuman MI, et al. Updates on pediatric sepsis. J Am Coll Emerg Physicians Open 2020 Oct;1(5):981-993 [FREE Full text] [doi: 10.1002/emp2.12173] [Medline: 33145549]
27. Kim F, Polin RA, Hooven TA. Neonatal sepsis. BMJ 2020 Oct 01;371:m3672. [doi: 10.1136/bmj.m3672] [Medline: 33004379]
28. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. Lancet 2017 Oct 14;390(10104):1770-1780. [doi: 10.1016/S0140-6736(17)31002-4] [Medline: 28434651]
29. Matics TJ, Sanchez-Pinto LN. Adaptation and validation of a pediatric sequential organ failure assessment score and evaluation of the sepsis-3 definitions in critically ill children. JAMA Pediatr 2017 Oct 02;171(10):e172352 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.2352] [Medline: 28783810]
30. Leclerc F, Duhamel A, Deken V, Grandbastien B, Leteurtre S, Groupe Francophone de Réanimation et Urgences Pédiatriques (GFRUP). Can the pediatric logistic organ dysfunction-2 score on day 1 be used in clinical criteria for sepsis in children? Pediatr Crit Care Med 2017 Aug;18(8):758-763. [doi: 10.1097/PCC.0000000000001182] [Medline: 28492402]
31. Schlapbach LJ, MacLaren G, Festa M, Alexander J, Erickson S, Beca J, et al. Prediction of pediatric sepsis mortality within 1 h of intensive care admission. Intensive Care Med 2017 Aug;43(8):1085-1096. [doi: 10.1007/s00134-017-4701-8] [Medline: 28220227]
32. Wynn JL, Polin RA. A neonatal sequential organ failure assessment score predicts mortality to late-onset sepsis in preterm very low birth weight infants. Pediatr Res 2020 Jul;88(1):85-90 [FREE Full text] [doi: 10.1038/s41390-019-0517-2] [Medline: 31394566]
33. Edwards SE, Grobman WA, Lappen JR, Winter C, Fox R, Lenguerrand E, et al. Modified obstetric early warning scoring systems (MOEWS): validating the diagnostic performance for severe sepsis in women with chorioamnionitis. Am J Obstet Gynecol 2015 Apr;212(4):536.e1-536.e8. [doi: 10.1016/j.ajog.2014.11.007] [Medline: 25446705]
34. Aarvold AB, Ryan HM, Magee LA, von Dadelszen P, Fjell C, Walley KR. Multiple organ dysfunction score is superior to the obstetric-specific sepsis in obstetrics score in predicting mortality in septic obstetric patients. Crit Care Med 2017 Jan;45(1):e49-e57 [FREE Full text] [doi: 10.1097/CCM.0000000000002018] [Medline: 27618276]
35. Bhattacharjee P, Edelson DP, Churpek MM. Identifying patients with sepsis on the hospital wards. Chest 2017 Apr;151(4):898-907 [FREE Full text] [doi: 10.1016/j.chest.2016.06.020] [Medline: 27374948]
36. Makam AN, Nguyen OK, Auerbach AD. Diagnostic accuracy and effectiveness of automated electronic sepsis alert systems: a systematic review. J Hosp Med 2015 Jun;10(6):396-402 [FREE Full text] [Medline: 25758641]
37. Wulff A, Montag S, Marschollek M, Jack T. Clinical decision-support systems for detection of systemic inflammatory response syndrome, sepsis, and septic shock in critically ill patients: a systematic review. Methods Inf Med 2019 Dec;58(S 02):e43-e57 [FREE Full text] [doi: 10.1055/s-0039-1695717] [Medline: 31499571]
38. Joshi M, Ashrafian H, Arora S, Khan S, Cooke G, Darzi A. Digital alerting and outcomes in patients with sepsis: systematic review and meta-analysis. J Med Internet Res 2019 Dec 20;21(12):e15166 [FREE Full text] [doi: 10.2196/15166] [Medline: 31859672]
39. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17 [FREE Full text] [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
40. Petersen C, Smith J, Freimuth RR, Goodman KW, Jackson GP, Kannry J, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. J Am Med Inform Assoc 2021 Mar 18;28(4):677-684 [FREE Full text] [doi: 10.1093/jamia/ocaa319] [Medline: 33447854]
41. Li L, Ackermann K, Baker J, Westbrook J. Use and evaluation of computerized clinical decision support systems for early detection of sepsis in hospitals: protocol for a scoping review. JMIR Res Protoc 2020 Nov 20;9(11):e24899 [FREE Full text] [doi: 10.2196/24899] [Medline: 33215998]
42. 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]
43. Schlapbach LJ. Paediatric sepsis. Curr Opin Infect Dis 2019 Oct;32(5):497-504. [doi: 10.1097/QCO.0000000000000583] [Medline: 31335441]

44. Ackermann K, Baker J, Green M, Fullick M, Varinli H, Westbrook J, et al. Computerized clinical decision support systems for the early detection of sepsis among adult inpatients: scoping review. J Med Internet Res 2022 Feb 23;24(2):e31083 [FREE Full text] [doi: 10.2196/31083] [Medline: 35195528]
45. Viswanathan M, Berkman N, Dryden D, Hartling L. Assessing risk of bias and confounding in observational studies of interventions or exposures: further development of the RTI item bank. In: AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US); 2013.
46. Ranganathan P, Aggarwal R. Study designs: part 1 - An overview and classification. Perspect Clin Res 2018;9(4):184-186 [FREE Full text] [Medline: 30319950]
47. Aggarwal R, Ranganathan P. Study designs: part 2 - descriptive studies. Perspect Clin Res 2019;10(1):34-36 [FREE Full text] [doi: 10.4103/picr.PICR_154_18] [Medline: 30834206]
48. Ranganathan P, Aggarwal R. Study designs: part 3 - Analytical observational studies. Perspect Clin Res 2019;10(2):91-94 [FREE Full text] [doi: 10.4103/picr.PICR_35_19] [Medline: 31008076]
49. Aggarwal R, Ranganathan P. Study designs: part 4 - Interventional studies. Perspect Clin Res 2019;10(3):137-139 [FREE Full text] [doi: 10.4103/picr.PICR_91_19] [Medline: 31404185]
50. Aggarwal R, Ranganathan P. Study designs: part 5 - interventional studies (II). Perspect Clin Res 2019;10(4):183-186 [FREE Full text] [doi: 10.4103/picr.PICR_138_19] [Medline: 31649869]
51. Ergonomics of human-system interaction - Part 11: usability: definitions and concepts. ISO standard no. 9241-11:2018(EN). The International Organization for Standardization. 2018. URL: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en [accessed 2021-02-26]
52. Achten NB, Dorigo-Zetsma JW, van der Linden PD, van Brakel M, Plötz FB. Sepsis calculator implementation reduces empiric antibiotics for suspected early-onset sepsis. Eur J Pediatr 2018 May;177(5):741-746. [doi: 10.1007/s00431-018-3113-2] [Medline: 29455368]
53. Achten NB, Visser DH, Tromp E, Groot W, van Goudoever JB, Plötz FB. Early onset sepsis calculator implementation is associated with reduced healthcare utilization and financial costs in late preterm and term newborns. Eur J Pediatr 2020 May;179(5):727-734 [FREE Full text] [doi: 10.1007/s00431-019-03510-9] [Medline: 31897840]
54. Arora V, Strunk D, Furqan SH, Schweig L, Lefaiver C, George J, et al. Optimizing antibiotic use for early onset sepsis: a tertiary NICU experience. J Neonatal Perinatal Med 2019;12(3):301-312. [doi: 10.3233/NPM-180075] [Medline: 30932898]
55. Balamuth F, Alpern ER, Abbadessa MK, Hayes K, Schast A, Lavelle J, et al. Improving recognition of pediatric severe sepsis in the emergency department: contributions of a vital sign-based electronic alert and bedside clinician identification. Ann Emerg Med 2017 Dec;70(6):759-68.e2 [FREE Full text] [doi: 10.1016/j.annemergmed.2017.03.019] [Medline: 28583403]
56. Beavers JB, Bai S, Perry J, Simpson J, Peeples S. Implementation and evaluation of the early-onset sepsis risk calculator in a high-risk university nursery. Clin Pediatr (Phila) 2018 Aug;57(9):1080-1085. [doi: 10.1177/0009922817751337] [Medline: 29284278]
57. Blumenthal E, Hooshvar N, Tancioco V, Newman R, Senderoff D, McNulty J. 238: Maternal Early Warning Trigger tool improves clinical care in large community hospital system. In: Proceedings of the SMFM 40th Annual Meeting--The Pregnancy Meeting. 2020 Presented at: SMFM 40th Annual Meeting--The Pregnancy Meeting; Feb 3-8, 2020; Grapevine, TX, United States URL: https://doi.org/10.1016/j.ajog.2019.11.254 [doi: 10.1016/j.ajog.2019.11.254]
58. Coffman Z, Smith J, Hubbard C, Chen E, McDaniel L. Using the electronic medical record to create a pediatric sepsis alert. In: Proceedings of the National Conference on Education. 2018 Presented at: National Conference on Education; Oct 22-25, 2016; San Francisco, CA, United States URL: https://publications.aap.org/pediatrics/article/141/1_MeetingAbstract/444/1975/Using-the-Electronic-Medical-Record-to-Create-a
59. Cruz AT, Williams EA, Graf JM, Perry AM, Harbin DE, Wuestner ER, et al. Test characteristics of an automated age- and temperature-adjusted tachycardia alert in pediatric septic shock. Pediatr Emerg Care 2012 Sep;28(9):889-894. [doi: 10.1097/PEC.0b013e318267a78a] [Medline: 22929140]
60. Davis T, Beigi R, Petticord V, Manetta L, Simhan H. Accurate identification of obstetric sepsis using e-record tools UPMC Magee-women's hospital. Pennsylvania Patient Safety Advisory 2018;15(supp. 1):53 [FREE Full text]
61. Dewan M, Vidrine R, Zackoff M, Paff Z, Seger B, Pfeiffer S, et al. Design, implementation, and validation of a pediatric ICU sepsis prediction tool as clinical decision support. Appl Clin Inform 2020 Mar;11(2):218-225 [FREE Full text] [doi: 10.1055/s-0040-1705107] [Medline: 32215893]
62. Dhudasia MB, Mukhopadhyay S, Puopolo KM. Implementation of the sepsis risk calculator at an academic birth hospital. Hosp Pediatr 2018 May;8(5):243-250. [doi: 10.1542/hpeds.2017-0180] [Medline: 29666161]
63. Eason J, Ward H, Danko O, Richardson K, Vaitkute R, McKeon-Carter R. Early-onset sepsis: can we screen fewer babies safely? Arch Dis Child 2021 Jan;106(1):86-88. [doi: 10.1136/archdischild-2019-317047] [Medline: 31678929]
64. Eisenberg M, Freiman E, Capraro A, Madden K, Monuteaux MC, Hudgins J, et al. Comparison of manual and automated sepsis screening tools in a pediatric emergency department. Pediatrics 2021 Feb;147(2):e2020022590. [doi: 10.1542/peds.2020-022590] [Medline: 33472987]
65. Emmanuel J, Torres A. The impact of automated electronic surveillance of electronic medical records on pediatric inpatient care. Cureus 2018 Oct 01;10(10):e3395 [FREE Full text] [doi: 10.7759/cureus.3395] [Medline: 30533330]

66. Fowler NT, Garcia M, Hankins C. Impact of integrating a neonatal early-onset sepsis risk calculator into the electronic health record. Pediatr Qual Saf 2019;4(6):e235 [FREE Full text] [doi: 10.1097/pq9.0000000000000235] [Medline: 32010861]
67. Gievers LL, Sedler J, Phillipi CA, Dukhovny D, Geddes J, Graven P, et al. Implementation of the sepsis risk score for chorioamnionitis-exposed newborns. J Perinatol 2018 Nov;38(11):1581-1587. [doi: 10.1038/s41372-018-0207-7] [Medline: 30158677]
68. Goyack L, Ma D, Miller J, Ye G, Mattox T, Torres A. 880: Evaluation of an automated electronic surveillance software program of an electronic medical record. In: Proceedings of the 45th Critical Care Congress of the Society of Critical Care Medicine, SCCM 2015. 2015 Presented at: 45th Critical Care Congress of the Society of Critical Care Medicine, SCCM 2015; Feb 20-24, 2016; Orlando, FL, United States URL: https://doi.org/10.1097/01.ccm.0000474708.59510.b5 [doi: 10.1097/01.ccm.0000474708.59510.b5]
69. Gur I, Riskin A, Markel G, Bader D, Nave Y, Barzilay B, et al. Pilot study of a new mathematical algorithm for early detection of late-onset sepsis in very low-birth-weight infants. Am J Perinatol 2015 Mar;32(4):321-330. [doi: 10.1055/s-0034-1384645] [Medline: 25077471]
70. Klingaman C, King L, Neff-Bulger M. Improved newborn care: evidence-based protocol for the evaluation and management of early-onset sepsis. Am J Med Qual 2018;33(1):106. [doi: 10.1177/1062860617741437] [Medline: 29139318]
71. Lloyd JK, Ahrens EA, Clark D, Dachenhaus T, Nuss KE. Automating a manual sepsis screening tool in a pediatric emergency department. Appl Clin Inform 2018 Oct;9(4):803-808 [FREE Full text] [doi: 10.1055/s-0038-1675211] [Medline: 30381818]
72. Mahdally S, Kim S. Implementing a model to predict risk of neonatal sepsis in late-preterm and term infants: a quality improvement initiative. In: Proceedings of the American Academy of Pediatrics National Conference and Exhibition 2017. 2018 Presented at: American Academy of Pediatrics National Conference and Exhibition 2017; Sep 16-19, 2017; Chicago, IL, United States URL: https://publications.aap.org/pediatrics/article/142/1_MeetingAbstract/193/2465/Implementing-a-Model-to-Predict-Risk-of-Neonatal
73. Mangubat PM, Shah S. 105 comparing accuracy of 2 phases of a pediatric electronic severe sepsis screening algorithm. In: Proceedings of the Critical Care Congress 2015. 2014 Presented at: Critical Care Congress 2015; Jan 17-21, 2015; Phoenix, AZ, United States URL: https://doi.org/10.1097/01.ccm.0000457602.28660.aa [doi: 10.1097/01.ccm.0000457602.28660.aa]
74. Salomon J, Serrao K, Guardiola J, Bonura A. 1335: Early detection of pediatric sepsis using an electronic medical record-based screening tool. In: Proceedings of the 46th Critical Care Congress of the Society of Critical Care Medicine, SCCM 2016. 2016 Presented at: 46th Critical Care Congress of the Society of Critical Care Medicine, SCCM 2016; Jan 21-25, 2017; Honolulu, HI, United States URL: https://doi.org/10.1097/01.ccm.0000510009.95039.f3 [doi: 10.1097/01.ccm.0000510009.95039.f3]
75. Sharma V, Adkisson C, Gupta K. Managing infants exposed to maternal chorioamnionitis by the use of early-onset sepsis calculator. Glob Pediatr Health 2019;6:2333794X19833711 [FREE Full text] [doi: 10.1177/2333794X19833711] [Medline: 31008151]
76. Skey D, Walters J, Surujdyal J. Limiting newborn antibiotic exposure through refined sepsis screening. In: Proceedings of the American Academy of Pediatrics National Conference and Exhibition 2017. 2018 Presented at: American Academy of Pediatrics National Conference and Exhibition 2017; Sep 16-19, 2017; Chicago, IL, United States URL: https://publications.aap.org/pediatrics/article/142/1_MeetingAbstract/574/2994/Limiting-Newborn-Antibiotic-Exposure-through
77. Stinson HR, Viteri S, Koetter P, Stevens E, Remillard K, Parlow R, et al. Early experience with a novel strategy for assessment of sepsis risk: the shock huddle. Pediatr Qual Saf 2019;4(4):e197 [FREE Full text] [doi: 10.1097/pq9.0000000000000197] [Medline: 31572898]
78. Stipelman CH, Smith ER, Diaz-Ochu M, Spackman J, Stoddard G, Kawamoto K, et al. Early-onset sepsis risk calculator integration into an electronic health record in the nursery. Pediatrics 2019 Aug;144(2):e20183464. [doi: 10.1542/peds.2018-3464] [Medline: 31278210]
79. Strunk T, Buchiboyina A, Sharp M, Nathan E, Doherty D, Patole S. Implementation of the neonatal sepsis calculator in an Australian tertiary perinatal centre. Neonatology 2018;113(4):379-382. [doi: 10.1159/000487298] [Medline: 29514161]
80. Torres A, Goyack L, Negron J, Miller J, Ye G, Lawless S. 821: Automated electronic surveillance of electronic medical records for shock in pediatric inpatients. In: Proceedings of the 45th Critical Care Congress of the Society of Critical Care Medicine, SCCM 2015. 2015 Presented at: 45th Critical Care Congress of the Society of Critical Care Medicine, SCCM 2015; Feb 20-24, 2016; Orlando, FL, United States p. 206-207 URL: https://doi.org/10.1097/01.ccm.0000474649.45155.26 [doi: 10.1097/01.ccm.0000474649.45155.26]
81. Vidrine R, Zackoff M, Paff Z, Seger B, Satterlee M, Buenaventura E, et al. Improving timely recognition and treatment of sepsis in the pediatric ICU. Jt Comm J Qual Patient Saf 2020 May;46(5):299-307. [doi: 10.1016/j.jcjq.2020.02.005] [Medline: 32201121]
82. Viteri S, Koetter P, Stinson H, Stevens E, Frizzola M. 1541: Comparison of a pediatric septic shock electronic screening tool and pediatric early warning score. In: Proceedings of the 47th Society of Critical Care Medicine Critical Care Congress, SCCM 2018. 2018 Presented at: 47th Society of Critical Care Medicine Critical Care Congress, SCCM 2018; Feb 25-28, 2018; San Antonio, TX, United States URL: https://doi.org/10.1097/01.ccm.0000529542.20350.1a [doi: 10.1097/01.ccm.0000529542.20350.1a]

83. West A, Hallman M, Giles K, Guynn A, May W, Shah S. 1540: Accuracy of detecting clinically relevant severe sepsis in children using a real-time EMR algorithm. In: Proceedings of the 47th Society of Critical Care Medicine Critical Care Congress, SCCM 2018. 2018 Presented at: 47th Society of Critical Care Medicine Critical Care Congress, SCCM 2018; Feb 25-28, 2018; San Antonio, TX, United States URL: https://doi.org/10.1097/01.ccm.0000529541.26188.44 [doi: 10.1097/01.ccm.0000529541.26188.44]
84. Zayek M, Bhat J, Bonner K, Blake M, Peevy K, Jha OP, et al. Implementation of a modified neonatal early-onset sepsis calculator in well-baby nursery: a quality improvement study. Pediatr Qual Saf 2020;5(4):e330 [FREE Full text] [doi: 10.1097/pq9.0000000000000330] [Medline: 32766501]
85. Lane RD, Funai T, Reeder R, Larsen GY. High reliability pediatric septic shock quality improvement initiative and decreasing mortality. Pediatrics 2016 Oct;138(4):e20154153. [doi: 10.1542/peds.2015-4153] [Medline: 27604184]
86. Larsen GY, Brilli R, Macias CG, Niedner M, Auletta JJ, Balamuth F, Improving Pediatric Sepsis Outcomes Collaborative Investigators. Development of a quality improvement learning collaborative to improve pediatric sepsis outcomes. Pediatrics 2021 Jan;147(1):e20201434 [FREE Full text] [doi: 10.1542/peds.2020-1434] [Medline: 33328337]
87. Escobar GJ, Puopolo KM, Wi S, Turk BJ, Kuzniewicz MW, Walsh EM, et al. Stratification of risk of early-onset sepsis in newborns ≥ 34 weeks' gestation. Pediatrics 2014 Jan;133(1):30-36 [FREE Full text] [doi: 10.1542/peds.2013-1689] [Medline: 24366992]
88. Puopolo KM, Draper D, Wi S, Newman TB, Zupancic J, Lieberman E, et al. Estimating the probability of neonatal early-onset infection on the basis of maternal risk factors. Pediatrics 2011 Nov;128(5):e1155-e1163 [FREE Full text] [doi: 10.1542/peds.2010-3464] [Medline: 22025590]
89. Neonatal early-onset sepsis calculator. Kaiser Permanente Division of Research. URL: https://neonatalsepsiscalculator.kaiserpermanente.org/ [accessed 2022-04-13]
90. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA 2016 Feb 23;315(8):762-774 [FREE Full text] [doi: 10.1001/jama.2016.0288] [Medline: 26903335]
91. Schlapbach LJ, Kissoon N. Defining pediatric sepsis. JAMA Pediatr 2018 Apr 01;172(4):312-314. [doi: 10.1001/jamapediatrics.2017.5208] [Medline: 29459982]
92. Scott HF, Deakyne SJ, Woods JM, Bajaj L. The prevalence and diagnostic utility of systemic inflammatory response syndrome vital signs in a pediatric emergency department. Acad Emerg Med 2015 Apr;22(4):381-389 [FREE Full text] [doi: 10.1111/acem.12610] [Medline: 25778743]
93. Weiss SL, Fitzgerald JC, Maffei FA, Kane JM, Rodriguez-Nunez A, Hsing DD, SPROUT Study InvestigatorsPediatric Acute Lung InjurySepsis Investigators Network. Discordant identification of pediatric severe sepsis by research and clinical definitions in the SPROUT international point prevalence study. Crit Care 2015 Sep 16;19:325 [FREE Full text] [doi: 10.1186/s13054-015-1055-x] [Medline: 26373923]
94. van Nassau SC, van Beek RH, Driessen GJ, Hazelzet JA, van Wering HM, Boeddha NP. Translating sepsis-3 criteria in children: prognostic accuracy of age-adjusted quick sofa score in children visiting the emergency department with suspected bacterial infection. Front Pediatr 2018;6:266 [FREE Full text] [doi: 10.3389/fped.2018.00266] [Medline: 30327759]
95. Schlapbach LJ, Straney L, Bellomo R, MacLaren G, Pilcher D. Prognostic accuracy of age-adapted SOFA, SIRS, PELOD-2, and qSOFA for in-hospital mortality among children with suspected infection admitted to the intensive care unit. Intensive Care Med 2018 Feb;44(2):179-188 [FREE Full text] [doi: 10.1007/s00134-017-5021-8] [Medline: 29256116]
96. Weiss SL, Deutschman CS. Are septic children really just "septic little adults"? Intensive Care Med 2018 Mar;44(3):392-394 [FREE Full text] [doi: 10.1007/s00134-017-5041-4] [Medline: 29356850]
97. Menon K, Schlapbach LJ, Akech S, Argent A, Biban P, Carrol ED, Pediatric Sepsis Definition Taskforce of the Society of Critical Care Medicine. Criteria for pediatric sepsis-a systematic review and meta-analysis by the pediatric sepsis definition taskforce. Crit Care Med 2022 Jan 01;50(1):21-36 [FREE Full text] [doi: 10.1097/CCM.0000000000005294] [Medline: 34612847]
98. Preterm birth. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/preterm-birth [accessed 2022-04-13]
99. Greenberg RG, Kandefer S, Do BT, Smith PB, Stoll BJ, Bell EF, Eunice Kennedy Shriver National Institute of Child HealthHuman Development Neonatal Research Network. Late-onset sepsis in extremely premature infants: 2000-2011. Pediatr Infect Dis J 2017 Aug;36(8):774-779 [FREE Full text] [doi: 10.1097/INF.0000000000001570] [Medline: 28709162]
100. Achten NB, Klingenberg C, Benitz WE, Stocker M, Schlapbach LJ, Giannoni E, et al. Association of use of the neonatal early-onset sepsis calculator with reduction in antibiotic therapy and safety: a systematic review and meta-analysis. JAMA Pediatr 2019 Nov 01;173(11):1032-1040 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.2825] [Medline: 31479103]
101. Pettinger KJ, Mayers K, McKechnie L, Phillips B. Sensitivity of the Kaiser Permanente early-onset sepsis calculator: a systematic review and meta-analysis. EClinicalMedicine 2020 Feb;19:100227 [FREE Full text] [doi: 10.1016/j.eclinm.2019.11.020] [Medline: 32140666]
102. Bekhof J, Reitsma JB, Kok JH, Van Straaten IH. Clinical signs to identify late-onset sepsis in preterm infants. Eur J Pediatr 2013 Apr;172(4):501-508. [doi: 10.1007/s00431-012-1910-6] [Medline: 23271492]

103. Molloy EJ, Wynn JL, Bliss J, Koenig JM, Keij FM, McGovern M, on behalf of the Infection, Inflammation, ImmunologyImmunisation (I4) section of the ESPR. Neonatal sepsis: need for consensus definition, collaboration and core outcomes. Pediatr Res 2020 Jul;88(1):2-4. [doi: 10.1038/s41390-020-0850-5] [Medline: 32193517]
104. Bauer ME, Bauer ST, Rajala B, MacEachern MP, Polley LS, Childers D, et al. Maternal physiologic parameters in relationship to systemic inflammatory response syndrome criteria: a systematic review and meta-analysis. Obstet Gynecol 2014 Sep;124(3):535-541. [doi: 10.1097/AOG.0000000000000423] [Medline: 25162253]
105. Miller K, Capan M, Weldon D, Noaiseh Y, Kowalski R, Kraft R, et al. The design of decisions: matching clinical decision support recommendations to Nielsen's design heuristics. Int J Med Inform 2018 Sep;117:19-25 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.05.008] [Medline: 30032961]
106. Miller K, Mosby D, Capan M, Kowalski R, Ratwani R, Noaiseh Y, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. J Am Med Inform Assoc 2018 May 01;25(5):585-592 [FREE Full text] [doi: 10.1093/jamia/ocx118] [Medline: 29126196]
107. Sagar K, Saha A. A systematic review of software usability studies. Int J Inf Technol 2017:1-24 [FREE Full text] [doi: 10.1007/s41870-017-0048-1]
108. Ellsworth MA, Dziadzko M, O'Horo JC, Farrell AM, Zhang J, Herasevich V. An appraisal of published usability evaluations of electronic health records via systematic review. J Am Med Inform Assoc 2017 Jan;24(1):218-226 [FREE Full text] [doi: 10.1093/jamia/ocw046] [Medline: 27107451]
109. Jankovic I, Chen JH. Clinical decision support and implications for the clinician burnout crisis. Yearb Med Inform 2020 Aug;29(1):145-154 [FREE Full text] [doi: 10.1055/s-0040-1701986] [Medline: 32823308]
110. The Lancet Respiratory Medicine. Crying wolf: the growing fatigue around sepsis alerts. Lancet Respir Med 2018 Mar;6(3):161. [doi: 10.1016/S2213-2600(18)30072-9] [Medline: 29508700]
111. Ernst FR, Levy H, Qualy RL. Simplified pharmacoeconomics of critical care and severe sepsis. J Intensive Care Med 2007;22(5):283-293. [doi: 10.1177/0885066607304231] [Medline: 17895486]
112. Arefian H, Heublein S, Scherag A, Brunkhorst FM, Younis MZ, Moerer O, et al. Hospital-related cost of sepsis: a systematic review. J Infect 2017 Feb;74(2):107-117. [doi: 10.1016/j.jinf.2016.11.006] [Medline: 27884733]
113. Hajj J, Blaine N, Salavaci J, Jacoby D. The "Centrality of sepsis": a review on incidence, mortality, and cost of care. Healthcare (Basel) 2018 Jul 30;6(3):90 [FREE Full text] [doi: 10.3390/healthcare6030090] [Medline: 30061497]

Abbreviations

CCDS: computerized clinician decision support
KPC: Kaiser Permanente early-onset sepsis risk calculator
LILACS: Latin American and Caribbean Health Sciences Literature
PQDT: ProQuest Dissertations and Theses Global
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews

Edited by C Lovis; submitted 02.12.21; peer-reviewed by M Dewan, M Sehgal, J Woods; comments to author 20.01.22; revised version received 27.02.22; accepted 19.03.22; published 06.05.22.

Please cite as:
Ackermann K, Baker J, Festa M, McMullan B, Westbrook J, Li L
Computerized Clinical Decision Support Systems for the Early Detection of Sepsis Among Pediatric, Neonatal, and Maternal Inpatients: Scoping Review
JMIR Med Inform 2022;10(5):e35061
URL: https://medinform.jmir.org/2022/5/e35061
doi:10.2196/35061
PMID:35522467

Review

Comparison of Severity of Illness Scores and Artificial Intelligence Models That Are Predictive of Intensive Care Unit Mortality: Meta-analysis and Review of the Literature

Cristina Barboi[1], MSc, MD; Andreas Tzavelis[2,3], BSc, BA; Lutfiyya NaQiyba Muhammad[4], MPH, PhD

[1]Indiana University Purdue University, Regenstrief Institue, Indianapolis, IN, United States

[2]Medical Scientist Training Program, Feinberg School of Medicine, Chicago, IL, United States

[3]Department of Biomedical Engineering, Northwestern University, Chicago, IL, United States

[4]Department of Preventive Medicine and Biostatistics, Northwestern University, Evanston, IL, United States

Corresponding Author:
Cristina Barboi, MSc, MD
Indiana University Purdue University
Regenstrief Institue
1101 W 10th St,
Indianapolis, IN, 46202
United States
Phone: 1 2628538872
Email: cbarboi@iu.edu

Abstract

Background: Severity of illness scores—Acute Physiology and Chronic Health Evaluation, Simplified Acute Physiology Score, and Sequential Organ Failure Assessment—are current risk stratification and mortality prediction tools used in intensive care units (ICUs) worldwide. Developers of artificial intelligence or machine learning (ML) models predictive of ICU mortality use the severity of illness scores as a reference point when reporting the performance of these computational constructs.

Objective: This study aimed to perform a literature review and meta-analysis of articles that compared binary classification ML models with the severity of illness scores that predict ICU mortality and determine which models have superior performance. This review intends to provide actionable guidance to clinicians on the performance and validity of ML models in supporting clinical decision-making compared with the severity of illness score models.

Methods: Between December 15 and 18, 2020, we conducted a systematic search of PubMed, Scopus, Embase, and IEEE databases and reviewed studies published between 2000 and 2020 that compared the performance of binary ML models predictive of ICU mortality with the performance of severity of illness score models on the same data sets. We assessed the studies' characteristics, synthesized the results, meta-analyzed the discriminative performance of the ML and severity of illness score models, and performed tests of heterogeneity within and among studies.

Results: We screened 461 abstracts, of which we assessed the full text of 66 (14.3%) articles. We included in the review 20 (4.3%) studies that developed 47 ML models based on 7 types of algorithms and compared them with 3 types of the severity of illness score models. Of the 20 studies, 4 (20%) were found to have a low risk of bias and applicability in model development, 7 (35%) performed external validation, 9 (45%) reported on calibration, 12 (60%) reported on classification measures, and 4 (20%) addressed explainability. The discriminative performance of the ML-based models, which was reported as AUROC, ranged between 0.728 and 0.99 and between 0.58 and 0.86 for the severity of illness score–based models. We noted substantial heterogeneity among the reported models and considerable variation among the AUROC estimates for both ML and severity of illness score model types.

Conclusions: ML-based models can accurately predict ICU mortality as an alternative to traditional scoring models. Although the range of performance of the ML models is superior to that of the severity of illness score models, the results cannot be generalized due to the high degree of heterogeneity. When presented with the option of choosing between severity of illness score or ML models for decision support, clinicians should select models that have been externally validated, tested in the practice environment, and updated to the patient population and practice environment.

Trial Registration: PROSPERO CRD42021203871; https://tinyurl.com/28v2nch8

(JMIR Med Inform 2022;10(5):e35293) doi:10.2196/35293

KEYWORDS

artificial intelligence; machine learning; intensive care unit mortality; severity of illness models

Introduction

Background

In the United States, intensive care unit (ICU) care costs account for 1% of the US gross domestic product, underscoring the need to optimize its use to attenuate the continued increase in health care expenditures [1]. Models that characterize the severity of illnesses of patients who are critically ill by predicting complications and ICU mortality risk can guide organizational resource management and planning, implementation and support of critical clinical protocols, and benchmarking and are proxies for resource allocation and clinical performance [2]. Although the medical community values the information provided by such models, they are not consistently used in practice because of their complexity, marginal predictive capacity, and limited internal or external validation [2-5].

Severity of illness score models require periodic updates and customizations to reflect changes in medical care and regional case pathology [6]. Scoring models are prone to high interrater variability, are less accurate for patients with increased severity of illness score or specific clinical subgroups, are not designed for repeated applications, and cannot represent patients' status trends [7]. The Acute Physiology and Chronic Health Evaluation (APACHE)-II (APACHE-II) and Simplified Acute Physiology Score (SAPS), developed in the 80s, are still in use [8]. The underlying algorithms for APACHE-IV are in the public domain and are available at no cost; however, their use is time intensive and is facilitated by software that requires payments for licensing implementation and maintenance [9]. Compared with SAPS-III, which uses data exclusively obtained within the first hour of ICU admission [10], APACHE-IV uses data from the first day (24 hours) [11]. Although the Sequential Organ Failure Assessment (SOFA) is an organ dysfunction score that detects differences in the severity of illness and is not designed to predict mortality, it is currently used to estimate mortality risk based on the mean, highest, and time changes accrued in the score during the ICU stay [11].

The availability of machine-readable data from electronic health records enables the analysis of large volumes of medical data using machine learning (ML) methods. ML algorithms enable the exploration of high-dimensional data and the extraction of features to develop models that solve classification or regression problems. These algorithms can fit linear and nonlinear associations and interactions between predictive variables and relate all or some of the predictive variables to an outcome. The increased flexibility of ML models comes with the risk of overfitting training data; therefore, model testing on external data is essential to ensure adequate performance on previously unseen data. In model development, the balance between the model's accuracy and generalizability, or bias and variance, is achieved through model training on a *training set* and hyperparameter optimization on a *tuning set*. Once a few models have been trained, they can be internally validated on a *split-sample* data set or cross-validated; the candidate model chosen is then validated on an unseen *test data set* to calculate its performance metrics and out of sample error [12]. The choice of algorithm is critical for providing a balance between interpretability, accuracy, and susceptibility to bias and variance [13]. Compared with the severity of illness scores, ML models can incorporate large numbers of covariates and temporal data, nonlinear predictors, trends in measured variables, and complex interactions between variables [14]. Numerous ML algorithms have been integrated into ICU predictive models, such as artificial neural networks (NNs), deep reinforcement learning, support vector machines (SVMs), random forest models, genetic algorithms, clinical trajectory models, gradient boosting models, k-nearest neighbor, naive Bayes, and the Ensemble approach [15]. Despite the rapidly growing interest in using ML methods to support clinical care, modeling processes and data sources have been inadequately described [16,17]. Consequently, the ability to validate and generalize the current literature's results is questionable.

Objectives

This study aims to systematically review and meta-analyze studies that compare binary classification ML models with the severity of illness scores for predicting ICU mortality and determine which models have superior performance. This review intends to provide actionable guidance to clinicians on the prognostic value of ML models compared with the severity of illness scores in supporting clinical decision-making, as well as on their performance, in the context of the current guidelines [18] and recommendations for reporting ML analysis in clinical research [19] (Table 1).

Table 1. Recommended structure for reporting ML[a] models.

Research question and ML justification	Data sources and preprocessing (feature selection)	Model training and validation
Clinical question	Population	Hardware, software, and packages used
Intended use of the result	Sample record and measurement characteristics	Evaluation (calibration and discrimination)
Defined problem type	Data collection and quality	Configuration (parameters and hyperparameters)
Available data	Data structure and types	Model optimization and generalization (hyperparameter tuning and parameter limits)
Defined ML method and rationale	Differences between evaluation and validation sets	Validation method and data split and cross-validation
Defined evaluation measures, training protocols, and validation	Data preprocessing (data aggregation, missing data, transformation, and label source)	Validation method performance metrics on an external data set
N/A[b]	Input configuration	Reproducibility, code reuse, and explainability

[a]ML: machine learning.

[b]N/A: not applicable.

Methods

We conducted a systematic review of the relevant literature. The research methods and reporting followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement and guide to review and meta-analysis of prediction models [20,21].

Information Sources and Search Strategy

Between December 15 and 18, 2020, we performed a comprehensive search in the bibliographic databases PubMed, Scopus, Embase, and IEEE of the literature published between December 2000 and December 15, 2020. These databases were available free of charge from the university library. We selected PubMed for its significance in biomedical electronic research; Scopus for its wide journal range, keyword search, and citation analysis; Embase because of its European union literature coverage; and IEEE Xplore for its access to engineering and computer science literature.

The search terms included control terms (Medical Subject Headings and Emtree) and free-text terms. The filters applied during the search of all 4 databases were *Humans* and *Age:Adult*. A search of the PubMed database using the terms (*AI artificial intelligence*) OR (*machine learning*) AND (*intensive care unit*) AND (*mortality)* identified 125 articles. The Scopus database was searched using the terms KEY (*machine learning*) OR KEY (*artificial-intelligence*) AND KEY (*intensive care unit*) AND KEY (*mortality*) revealed 182 articles. The Embase database queries using the terms (*AI Artificial Intelligence*) OR (*machine learning*) AND (*intensive care unit*) AND (*mortality*) resulted in 103 articles. The IEEE database search using the terms (*machine learning*) OR (*artificial intelligence*) AND (*intensive care unit*) AND (*mortality*) produced 51 citations.

A total of 2 authors (CB and AT) screened titles and abstracts and recorded the reasons for exclusion. The same authors (CB and AT) independently reviewed the previously selected full-text articles to determine their eligibility for quantitative and qualitative assessments. Both authors revisited the discrepancies to guarantee database accuracy and checked the references of the identified articles for additional papers. A third researcher (LNM) was available to resolve any disagreements.

Eligibility Criteria and Study Selection

We included studies that compared the predictive performance of newly developed ML classification models predictive of ICU mortality with the severity of illness score models on the same data sets in the adult population. To be included in the review, the studies had to provide information on the patient cohort, model development and validation, and performance metrics. Both prospective and retrospective studies were eligible for inclusion.

Data Collection Process

Data extraction was performed by CB, reviewed by AT, and guided by the CHARMS (Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies) checklist [22] specifically designed for systematic reviews of prognostic prediction models. The methodological qualities of the included studies were appraised with guidance from the Prediction model Risk of Bias (ROB) Assessment Tool (PROBAST) [23]. The reported features of the ML models are shown in Table 2.

Table 2. CHARMS (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies) checklist.

Author	Data source (description)	Outcome mortality	Data preparation		Model training			Predictive performance				Generalizability			
			A[a]	B[b]	C[c]	D[d]	E[e]	F[f]	G[g]	H[h]	I[i]	J[j]	K[k]	L[l]	M[m]	N[n]
Pirracchio et al [1]	MIMIC[o] 2	Hospital				✓	✓	✓	✓			✓		✓		✓
Nielsen et al [24]	Danish ICU[p]	Hospital 30/90 days	✓	✓	✓	✓	✓		✓	✓		✓			✓	
Nimgaonkar et al [25]	ICU India	Hospital				✓	✓	✓	✓							
Xia et al [26]	MIMIC 3	28 days/hospital	✓	✓		✓	✓		✓	✓					✓	✓
Purushotham et al [27]	MIMIC 3	Hospital, 2 days, 3 days, 30 days, 1 year	✓	✓		✓	✓		✓	✓		✓			✓	✓
Nanayakkara et al [28]	ANZICS[q]	Hospital	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓
Meyer et al [29]	Germany	Hospital		✓		✓	✓		✓	✓		✓		✓		✓
Meiring et al [7]	CCHIC[r] United Kingdom	Hospital	✓	✓	✓	✓	✓		✓						✓	✓
Lin et al [30]	MIMIC 3	Hospital	✓	✓		✓	✓	✓	✓	✓						
Krishnan et al [31]	MIMIC 3	ICU	✓	✓		✓	✓		✓	✓				✓		
Kang et al [32]	Korea	Hospital	✓		✓	✓	✓	✓	✓	✓			✓			
Johnson et al [33]	United Kingdom	ICU and hospital	✓	✓		✓	✓	✓	✓		✓	✓		✓		
Holmgren et al [34]	Sweden	Hospital and 30 days			✓	✓	✓	✓	✓	✓					✓	✓
Garcia-Gallo et al [35]	MIMIC 3	Hospital and 1 year	✓	✓	✓	✓	✓	✓	✓	✓					✓	✓
El-Rashidy et al [36]	MIMIC 3	ICU and hospital	✓	✓		✓	✓		✓	✓		✓			✓	✓
Silva et al [37]	EURICUS[s] 2	ICU	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓		
Caicedo-Torres et al [38]	MIMIC 3	ICU	✓	✓		✓	✓		✓	✓			✓			
Deshmukh et al [39]	eICU-CRD[t]	ICU	✓	✓	✓	✓	✓		✓	✓			✓		✓	
Ryan et al [40]	MIMIC 2	ICU and hospital	✓	✓	✓	✓	✓		✓	✓		✓			✓	✓
Mayaud et al [41]	MIMIC 2	Hospital	✓	✓		✓	✓	✓	✓		✓					

[a]Data normalization/outlier addressed.

[b]Missing data addressed.

[c]Hyperparameter optimization addressed.

[d]Overfitting/shrinkage and cross-validation addressed.

[e]Predictor selection, full model versus backward elimination.

[f]Calibration assessed (Brier, Hosmer-Lemeshow, and calibration plot).

[g]Discrimination/reclassification performed (net reclassification improvement/integrated discrimination improvement).

[h]Classification reported.

[i]Recalibration performed.

[j]Externally validated.

[k]Explainability addressed/decision curve analysis.
[l]Clinical applicability addressed.
[m]Prediction span defined.
[n]Intended moment of use reported.
[o]MIMIC: Medical Information Mart for Intensive Care.
[p]ICU: intensive care unit.
[q]ANZICS: Australia New Zealand Intensive Care Unit Society.
[r]CCHIC: Critical Care Health Informatics Collaborative.
[s]EURICUS: European ICU studies.
[t]eICU CRD: Electronic ICU Collaborative Research Database.

Assessment of the ROB and Quality of Reviewed Studies

The reviewers used the PROBAST tool to assess the methodological quality of each study for ROB and concerns regarding applicability in 4 domains: study participants, predictors, outcome, and analysis [23]. The reviewers evaluated the applicability of the selected studies by assessing the extent to which the studied outcomes matched the goals of the review in the 4 domains. We evaluated the ROB by assessing the primary study design and conduct, predictor selection process, outcome definition, and performance analysis. The ROB in the reporting models' performance was appraised by exploring the reported measures of calibration (model's predicted risk of mortality vs the observed risk), discrimination (model's ability to discriminate between patients who are alive or expired), classification (sensitivity and specificity), and reclassification (net reclassification index). The performance of the models on internal data sets not used for model development—internal validation—and on data sets originating from an external patient population–external validation—were weighted in the ROB assignment. The ROB and applicability were assigned as *low risk*, *high risk*, or *unclear risk* according to PROBAST recommendations [42].

Meta-analysis and Performance Metrics

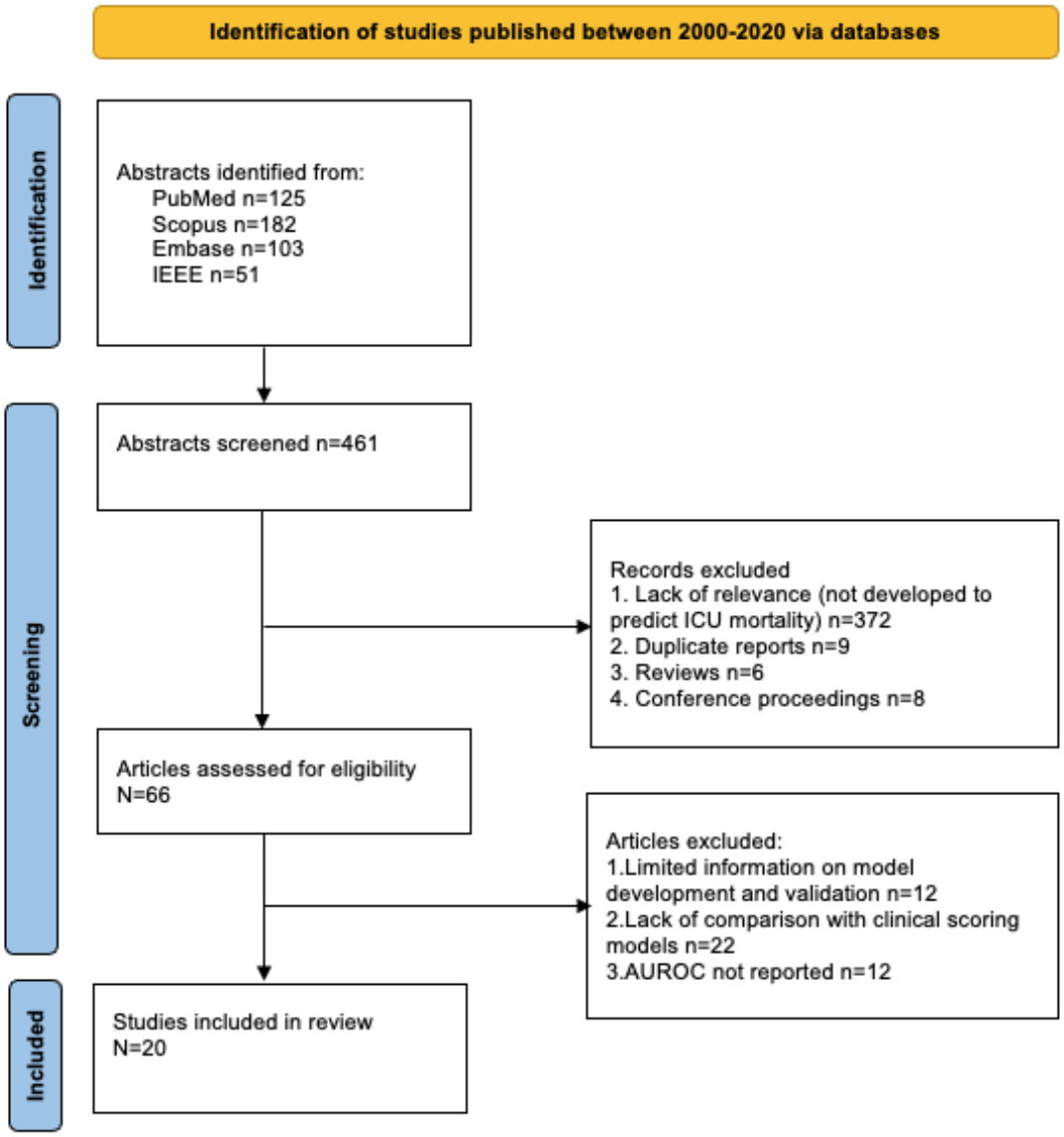
The C statistic–area under the receiver operating curve (AUROC) is the most commonly reported estimate of discriminative performance for binary outcomes [43-46] and the pragmatic performance measure of ML and severity of illness score models previously used in the medical literature to compare models based on different computational methods [21,45-47]. It is generally interpreted as follows: an AUROC of 0.5 suggests no discrimination, 0.7 to 0.8 is considered acceptable performance, 0.8 to 0.9 is considered excellent performance, and >0.9 is considered outstanding performance [48]. We included the performance of models developed using similar algorithms in forest plots and performed heterogeneity diagnostics and investigations without calculating a pooled estimate [49]. The results were pooled only for studies that followed a consistent methodology that included the external validation or benchmarking of the models. Random-effects meta-analyses computed the pooled AUROC for the following subgroups of ML algorithms—NNs and Ensemble—and the following subgroups of scoring models—SAPS II, APACHE II and SOFA. The AUROC for each model type was weighted using the inverse of its variance. Pooled AUROC estimates for each model were meta-analyzed along with 95% CIs of the estimates and were reported in forest plots together with the associated heterogeneity statistics (I^2, τ^2, and Cochran Q). Cochran Q statistic (also known as the chi-square statistic) determines the within-study variation, τ^2 determines the between-study variability, and I^2 represents the percentage of variability from the AUROC estimate not caused by sampling error [36]. The Cochran Q *P* value is denoted as *P*. Meta-analyses were conducted in R (version 3.6.1) [37] (see Multimedia Appendix 1 for scripts).

Results

Selection Process

Of the 461 screened abstracts, we excluded 372 (80.7%) because of relevance (models not developed to predict ICU mortality), 9 (2%) duplicates, 6 (1.3%) reviews, and 8 (1.7%) conference proceedings (not intended for clinical application). We assessed the full text of 66 articles; the most common performance method reported to allow comparison between all models and a meta-analysis was the C statistic–AUROC. Of the 66 articles, we excluded 12 (18%) articles because of limited information on model development, 22 (33%) articles because of a lack of comparison with clinical scoring models, and 12 (18%) articles as the AUROC was not reported. The search strategy and selection process are illustrated in Figure 1.

Figure 1. Search strategy and selection process. AUROC: area under the receiver operating curve; ICU: intensive care unit.

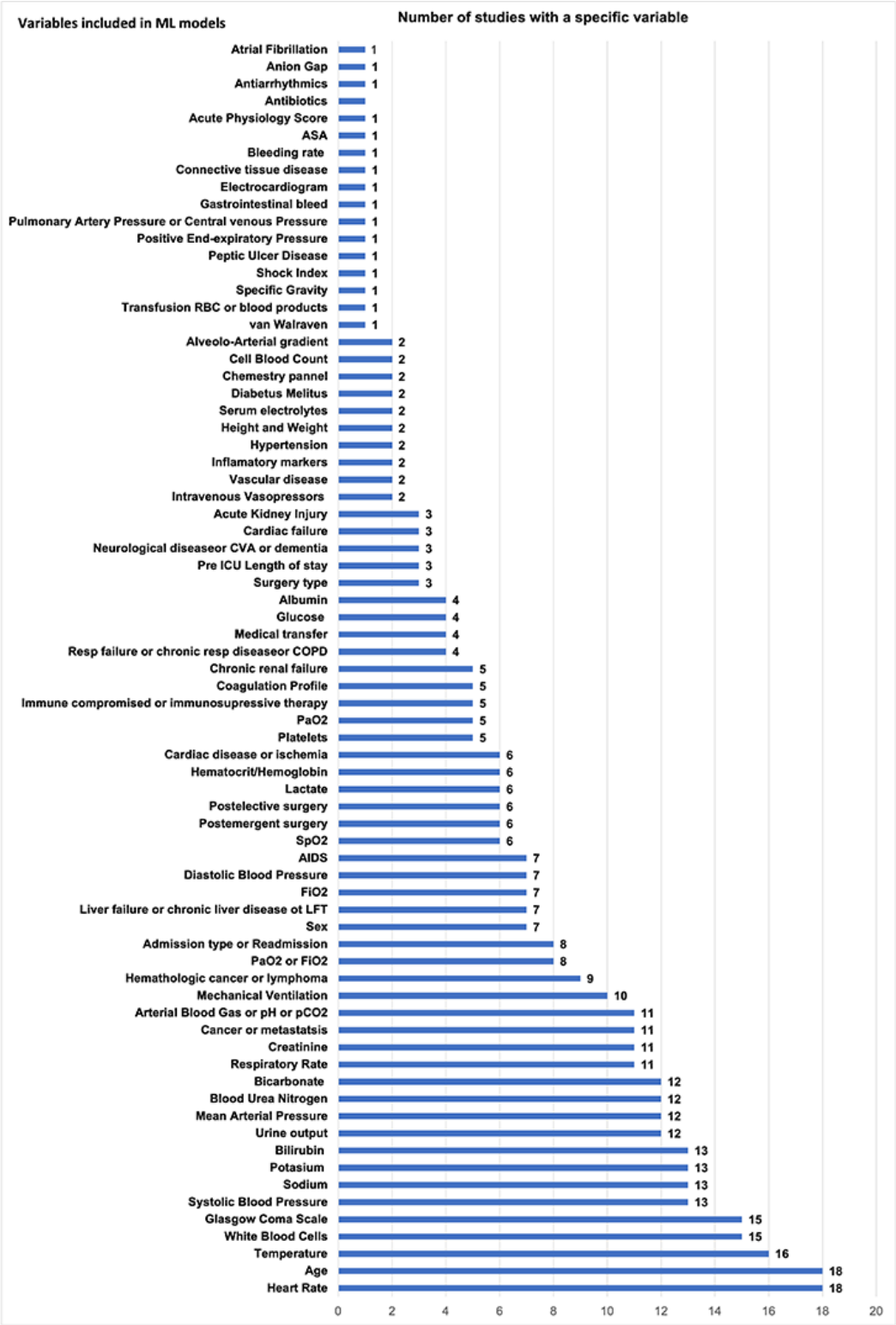


Assessment of the Prediction Model Development

The 20 studies reported 47 ML models that were developed based on 7 types of algorithms and compared them with 3 severity of illness score models. All ML models were developed through a retrospective analysis of the ICU data sets. Of the 20 studies, 10 (50%) used data from the publicly available Medical Information Mart for Intensive Care database (Beth Israel Deaconess Medical Center in the United States) at different stages of expansion. Of the 20 studies, 10 (50%) used national health care databases (Danish, Australia-New Zealand, United Kingdom, and Sweden) or ICU-linked databases (Korea, India, and the United Kingdom). One of the studies included data from >80 ICUs belonging to >40 hospitals [33], and one of the studies' ICU-linked database collected data from 9 European countries [37]. The cohorts generating the data sets used for model development and internal testing ranged from 1571 to 217,289 patients, with a median of 15,789 patients. Of the 20 studies, 10 (50%) used data from patients admitted to general ICUs, whereas 10 (50%) studies used data from patients who were critically ill with specific pathologies: gastrointestinal bleeds [39], COVID-19 and pneumonia–associated respiratory failure [40], postcardiac arrest [28], postcardiac surgery [29,36], acute renal insufficiency [30,32], sepsis [35,41], or neurological pathology [25]. The lower age thresholds for study inclusion ranges were 12 years [25], 15 years [26,27], 16 years [33,35,38], 18 years [24,29,40], and 19 years [30]. Within the studied cohorts, mortality ranged from 0.08 to 0.5 [29,32,36].

The processes and tools used for the selection of predicting variables were described in 65% (13/20) of studies and included the least absolute shrinkage and selection operator, stochastic gradient boosting [33,35], genetic algorithms, and particle swarm optimization [33]. Approximately 15% (3/20) of studies [25,26,35] reported multiple models developed on variable predictor sets, which were subsequently tested for the best performance, validation, and calibration. The number of predictive variables used in the final models varied between 1 and 80, with a median of 21. The most common predicting variables are shown in Figure 2 and are grouped by the frequency of occurrence in the studies.

Figure 2. Frequency and type of ML model input variables (x-axis: number of studies using the input variables; y-axis: input variable). ASA: American Society of Anesthesiology; COPD: chronic obstructive pulmonary disease; CVA: cerebral vascular accident; FIO2: fraction of inspired oxygen; ICU: intensive care unit; LFT: liver function test; ML: machine learning; RBC: red blood cell; SpO2: oxygen saturation; PaO2: arterial oxygen pressure; PaCO2: arterial CO2 pressure.



All studies developed models on 24-hour data; furthermore, ML models were developed on the first hour of ICU data [34]; the first 48-hour data [27,38,41]; 3-day data [40]; 5-day data [7]; 10-day data [26]; or on patients' prior medical history collected from 1 month, 3 months, 6 months, 1 year, 2.5 years, 5 years, 7.5 years, 10 years, and 23 years [24]. The frequency of data collection ranged from every 30 minutes [29], 1 hour [1,25,27,37], 3 hours, 6 hours, 12 hours, 15 hours [38], and 24 hours [7,36] to every 27 hours, 51 hours, and 75 hours [40].

Researchers handled missing data and continuous and fixed variables differently. A total of 6 model developers provided no information on missing data [1,25,29,32,34,39], and 1 [27] addressed data cleaning. Researchers [24,30,33,35-37,40,41]

removed the records with missing values ranging from 1 missing value per admission to 30%, 50%, and 60% of missing data. One of the studies [29] included only variables documented for at least 50% of the patients and imputed the missing values with the last measured value for the feature. Missing values (up to 60%) were forward-filled; backward-filled; or replaced with means (continuous variables) or modes (categorical variables), normal values, averages [24,28,36,38,40], predictive mean matching [7], or linear interpolation imputation method [26]. The data were normalized using the minimum-maximum normalization technique. The time prediction of hospital mortality was undefined in 45% (9/20) of studies and varied from 2 or 3 days to 28 days, 30 days [26], 90 days [24], and up to 1 year [24] in the others.

There was a wide range in the prevalence of mortality among studies (0.08-0.56), creating a class imbalance in the data sets. In studies with low investigated outcome mortality, few researchers addressed the problem of class imbalance (survivors vs nonsurvivors) through balanced training [24,37], random resampling [29], undersampling [36], or class penalty and reweighting schemes [38]. A breakdown of the model characteristics is presented in Table 3.

Table 3. Information on the ML[a] prediction model development, validation, and performance, and on the severity of illness score performance.

Author	ML model type (AUROC[b] test)	Data training/test (split %)	Features	K-fold/validation	External validation data set	ML AUROC external	Severity of illness score model type (AUROC)
Pirracchio et al, 2015 [1]	• Ensemble SICULA[c] (0.85)	24,508	17	5-fold cross-validation	200	0.94	• SAPS[d]-II (0.78) • APACHE[e]-II (0.83) • SOF[f] (0.71)
Nielsen et al, 2019 [24]	• NN[g] (0.792)	10,368 (80/20)	44	5-fold cross-validation	1528	0.773	• SAPS-II (0.74) • APACHE-II (0.72)
Nimgaonkar et al, 2004 [25]	• NN (0.88)	2962 (70/30)	15	N/A[h]	N/A	N/A	• APACHE-II (0.77)
Xia et al, 2019 [26]	• Ensemble-LSTM[i] (0.85) • LSTM (0.83) • DT[j] (0.82)	18,415 (90/10)	50	Bootstrap and RSM[k]	N/A	N/A	• SAPS-II (0.77) • SOFA (0.73) • APACHE-II (0.74)
Purushotham et al, 2018 [27]	• NN (0.87) • Ensemble (0.84)	35,627	17/22/136	5-fold cross-validation	External benchmark	N/A	• SAPS-II (0.80) • SOFA (0.73)
Nanayakkara et al, 2018 [28]	• DT (0.86) • SVM[l] (0.86) • NN (0.85) • Ensemble (0.87) • GBM[m] (0.87)	39,560 (90/10)	29	5-fold cross-validation	N/A	N/A	• APACHE-III (0.8)
Meyer et al, 2018 [29]	• NN (0.95)	5898 (90/10)	52	10-fold cross-validation	5989	0.81	• SAPS-II (0.71)
Meiring et al, 2018 [7]	• DT (0.85) • NN (0.86) • SVM (0.86)	80/20	25	21,911 LOO[n]	N/A	N/A	• APACHE-II (0.83)
Lin et al, 2019 [30]	• DT (0.86) • NN (0.83) • SVM (0.86)	19,044	15	5-fold cross-validation	N/A	N/A	• SAPS-II (0.79)
Krishnan et al, 2018 [31]	• NN-ELM[o] (0.99)	10,155 (75/25)	1	10-fold cross-validation	N/A	N/A	• SAPS (0.80) • SOFA (0.73) • APS[p]-III (0.79)
Kang et al, 2020 [32]	• SVM (0.77) • DT (0.78) • NN (0.776) • k-NN[q] (0.76)	1571 (70/30)	33	10-fold cross-validation	N/A	N/A	• SOFA (0.66) • APACHE-II (0.59)
Johnson et al, 2013 [33]	• LR[r] univariate (0.902) • LR multivariate (0.876)	39,070 (80/20)	10	10-fold cross-validation	23,618	0.837 (univariate); 0.868 (multivariate)	• APS-III (0.86)
Holmgren et al, 2019 [34]	• NN (0.89)	217,289 (80/20)	8	5-fold cross-validation	N/A	N/A	• SAPS-III (0.85)
Garcia-Gallo et al, 2020 [35]	• SGB-LASSO[s] (0.803)	5650 (70/30)	18 140 37	10-fold cross-validation	N/A	N/A	• SOFA (0.58) • SAPS (0.70)
El-Rashidy et al, 2020 [36]	• Ensemble (0.93)	10,664 (75/25)	80	10-fold cross-validation	External benchmark	N/A	• APACHE-II (0.73) • SAPS-II (0.81) • SOFA-II (0.78)

Author	ML model type (AUROC[b] test)	Data training/test (split %)	Features	K-fold/validation	External validation data set	ML AUROC external	Severity of illness score model type (AUROC)
Silva et al, 2006 [37]	• NN (0.85)	13,164 (66/33)	12	Hold out	N/A	N/A	• SAPS- II (0.8)
Caicedo-Torres et al, 2019 [38]	• NN (0.87)	22,413	22	5-fold cross-validation	N/A	N/A	• SAPS-II (0.73)
Deshmukh et al, 2020 [39]	• XGB[t] (0.85)	5691 (80/20)	34	5-fold cross-validation	N/A	N/A	• APACHE-IV (0.8)
Ryan et al, 2020 [40]	• DT (0.86)	35,061 (80/20)	12	5-fold cross-validation	114	0.91	• qSOFA[u] (0.76)
Mayaud et al, 2013 [41]	• GA[v]+LR (0.82)	2113 (70/30)	25	BBCCV[w]	N/A	N/A	• APACHE-III (0.68)

[a]ML: machine learning.
[b]AUROC: area under the receiver operating curve.
[c]SICULA: Super ICU Learner Algorithm.
[d]SAPS: Simplified Acute Physiology Score.
[e]APACHE: Acute Physiology and Chronic Health Evaluation.
[f]SOFA: Sequential Organ Failure Assessment.
[g]NN: neural network.
[h]N/A: not applicable.
[i]LSTM: long short-term memory.
[j]DT: decision tree.
[k]RSM: random subspace method.
[l]SVM: support vector machine.
[m]GBM: gradient boosting machine.
[n]LOO: leave one out.
[o]ELM: extreme learning machine.
[p]APS: Acute Physiology Score.
[q]k-NN: k-nearest neighbor.
[r]LR: logistic regression.
[s]SGB-LASSO: stochastic gradient boosting least absolute shrinkage and selection operator.
[t]XGB: extreme gradient boosting.
[u]qSOFA: Quick Sequential Organ Failure Assessment.
[v]GA: genetic algorithm.
[w]BBCV: bootstrap bias–corrected cross-validation.

Overview of ML Algorithms and Model Validation

The reviewers recorded the ML model types based on the final trained model structure rather than on the algorithm used for fitting the model (Table 3). The reviewers noted a diversity of strategies in model fitting, although the implemented models defined the operating functions and transformations. Of the 20 studies, NNs were applied in 13 (65%) [7,24-32,34,37,38], decision trees in 8 (40%) [7,26,28,30,32,35,39,40], SVM in 4 (20%) [7,28,30,32], and Ensemble of algorithms in 4 (20%) [1,27,28,36]. The types of algorithms used in the same study varied between 1 and 5. All studies provided information on data training and internal testing (see Table 2 for k-fold validation and data splitting). Of the 20 studies, 5 (25%) [1,24,29,33,40] performed validation on external data sets ranging from 114 to 23,618 patients, and 2 (10%) studies [27,36] benchmarked the ML model performance against existing ML mortality prediction models; 14 (70%) studies reported CIs for the measure of discrimination AUROC, 9 (45%) studies reported on calibration (Hosmer-Lemeshow, calibration curve, or Brier score), and 12 (60%) studies reported on classification measures (Table 4). Approximately 10% (2/20) of studies were available for use in clinical practice [1,33]; the models' decisions were explained with local interpretable model-agnostic explanations [28] or the Shapley additive explanations method (SHAP) [39]. The AUROC of the ML models ranged from 0.728 to 0.99 for predicting mortality.

Table 4. Reported performance measures of the ML[a] models.

Author and ML model	Classification measurements					Calibration measurements			Other
	Specificity	PPV[b]/precision	Recall/sensitivity	F_1 score	Accuracy	HL[c] score	Brier score	Calibration curve	
Pirrachio et al [1]									
Ensemble SL[d]-1	N/A[e]	N/A	N/A	N/A	N/A	N/A	0.079	U^f=0.0007 (calibration plot)	DS[g]=0.21
Ensemble SL-2	N/A	N/A	N/A	N/A	N/A	N/A	0.079	U=0.006 (calibration plot)	DS=0.26
Nielsen et al [24]									
NN[h]	N/A	0.388	N/A	N/A	N/A	N/A	N/A	N/A	Mathews correlation coefficient
Purushotham et al [27]									
NN	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.491 (AUPRC[i])
Ensemble	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.435 (AUPRC)
Nimgaonkar et al [25]									
NN-15 features	N/A	N/A	N/A	N/A	N/A	27.7	N/A	Calibration plot	N/A
NN-22 features	N/A	N/A	N/A	N/A	N/A	22.4	N/A	Calibration plot	N/A
Xia et al [26]									
Ensemble-LSTM[j]	0.7503	0.294	0.7758	0.4262	0.7533	N/A	N/A	N/A	N/A
LSTM	0.7746	0.305	0.7384	0.4317	0.7703	N/A	N/A	N/A	N/A
RF[k]	0.7807	0.306	0.71197	0.4290	0.7734	N/A	N/A	N/A	N/A
Nanayakkara et al [28]									
RF	0.79	0.75	0.76	N/A	0.78	N/A	0.156	Calibration plot	0.47 (log loss)
SVC[l]	0.81	0.77	0.75	N/A	0.78	N/A	0.153	Calibration plot	0.47 (log loss)
GBM[m]	0.78	0.75	0.8	N/A	0.79	N/A	0.147	Calibration plot	0.45 (log loss)
NN	0.72	0.71	0.82	N/A	0.77	N/A	0.158	Calibration plot	0.48 (log loss)
Ensemble	0.81	0.77	0.77	N/A	0.79	N/A	0.148	Calibration plot	0.45 (log loss)
Meyer et al [29]									
RNN[n]	0.91	0.9	0.85	0.88	0.88	N/A	N/A	N/A	N/A
Meiring et al [7]									
DT[o], NN, SVM[p]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Lin et al [30]									
RF	N/A	N/A	N/A	0.459	0.728	N/A	0.085	Calibration plot	N/A
NN	N/A	N/A	N/A	0.406	0.666	N/A	0.091	Calibration plot	N/A

Author and ML model	Classification measurements					Calibration measurements			Other
	Specificity	PPV[b]/precision	Recall/sensitivity	F_1 score	Accuracy	HL[c] score	Brier score	Calibration curve	
SVM	N/A	N/A	N/A	0.460	0.729	N/A	0.086	Calibration plot	N/A
Krishnan et al [31]									
ANN-ELM[q]	N/A	N/A	0.98	0.98	0.98	N/A	N/A	N/A	Mathews correlation coefficient
Kang et al [32]									
k-NN[r]	N/A	N/A	N/A	0.745	0.673	N/A	N/A	Calibration plot	N/A
SVM	N/A	N/A	N/A	0.752	0.696	N/A	N/A	Calibration plot	N/A
RF	N/A	N/A	N/A	0.762	0.69	N/A	N/A	Calibration plot	N/A
XGB[s]	N/A	N/A	N/A	0.763	0.711	N/A	N/A	Calibration plot	N/A
NN	N/A	N/A	N/A	0.749		N/A	N/A	Calibration plot	N/A
Johnson et al [33]									
LR[t] univariate	N/A	N/A	N/A	N/A	N/A	22	0.051	N/A	N/A
LR multivariate	N/A	N/A	N/A	N/A	N/A	19.6	0.048	N/A	N/A
Holmgren et al [34]									
NN	N/A	N/A	N/A	N/A	N/A	N/A	0.106	Calibration plot	N/A
Garcia-Gallo et al [35]									
SGB[u]	N/A	N/A	N/A	N/A	0.725	0.0916	N/A	Calibration plot	N/A
SGB-LASSO[v]	N/A	N/A	N/A	N/A	0.712	0.0916	N/A	Calibration plot	N/A
El-Rashidy et al [36]									
Ensemble	0.94	N/A	0.911	0.937	0.944	N/A	N/A	N/A	N/A
Silva et al [37]									
NN	0.79	N/A	0.78	N/A	0.7921	N/A	N/A	N/A	N/A
Caicedo-Torres et al [38]									
NN	0.827	N/A	0.75	N/A	N/A	N/A	N/A	N/A	N/A
Deshmukh et al [39]									
XGB	0.27	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A
Ryan et al [40]									
XGB	0.75	N/A	0.801	0.378	0.75	N/A	N/A	N/A	N/A
Mayaud et al [41]									

Author and ML model	Classification measurements					Calibration measurements			Other
	Specificity	PPV[b]/precision	Recall/sensitivity	F_1 score	Accuracy	HL[c] score	Brier score	Calibration curve	
GA[w]+LR	N/A	N/A	N/A	N/A	N/A	10.43	N/A	Calibration plot	N/A

[a]ML: machine learning.
[b]PPV: positive predictive value.
[c]HL: Hosmer-Lemeshow.
[d]SL: super learner.
[e]N/A: not available.
[f]*U* statistics.
[g]DS: discrimination slope.
[h]NN: neural network.
[i]AUPRC: area under the precison-recall curve.
[j]LSTM: long short-term memory.
[k]RF: random forest.
[l]SVC: support vector classifier.
[m]GBM: gradient boosting machine.
[n]RNN: recurrent neural network.
[o]DT: decision tree.
[p]SVM: support vector machine.
[q]ANN-ELM: artificial neural network extreme learning machine.
[r]k-NN: k-nearest neighbor.
[s]XGB: extreme gradient boosting.
[t]LR: logistic regression.
[u]SGB: stochastic gradient boosting.
[v]LASSO: least absolute shrinkage and selection operator.
[w]GA: genetic algorithm.

The performance of the ML models was compared with that of the following severity of illness scoring models: APACHE-II (6/20, 30%), APACHE-III (2/20, 10%), APACHE-IV (1/20, 5%), SAPS-II (11/20, 55%), SAPS-III (1/20, 5%), SOFA (9/20, 45%), and Acute Physiology Score-3 (2/20, 10%; Table 3). The severity of illness scores' discrimination reported as AUROC was associated with a CI in 65% (13/20) of studies. Calibration of the severity of illness score models was reported in 30% (6/20) of studies. Approximately 60% (12/20) of studies reported binary classification results. The severity of illness scores used for comparison and associated AUROCs were 0.70 to 0.803 for SAPS, 0.588 to 0.782 for SOFA, and 0.593 to 0.86 for APACHE (Table 3).

Analysis of ROB and Applicability

The results of the analysis of the ROB in the selection of the study population, predictors, outcome definition, and performance reporting are presented in Table 5. The results of the assessment of the developed ML models' applicability regarding the study participants and setting, the predictors used in the ML models' development and their timing, the outcome definition and prediction by the models, and the analysis that reports the models' performance are also presented in Table 5. Of the 47 models, 4 (9%) models [1,17,23,29] were identified as having a low risk, and 3 (6%) models were rated as having an uncertain ROB and applicability model development [24,29,40]. The main reason for the high ROB in the overall judgment of the study was the lack of external validation, which was identified in 28% (13/47) of the models.

Table 5. Assessment for ROB[a] and applicability for prognostic models with the Prediction model ROB Assessment Tool checklist.

Authors	ROB and applicability								
	Participants		Predictors		Outcome		Analysis	Overall judgment	
	ROB	Applicability	ROB	Applicability	ROB	Applicability	ROB	ROB	Applicability
Pirracchio et al [1]	Low[b]	Low	Low	Low	Low	Low	Low	Low	Low
Nielsen et al [24]	Low	Low	Unclear[c]	Low	Unclear	Low	Low	Unclear	Low
Nimgaonkar et al [25]	Low	Unclear	Low	Low	Low	Low	High[d]	High	Unclear
Xia et al [26]	Low	Low	Low	Low	Unclear	Low	High	High	Low
Purushotham et al [27]	Low	Low	Low	Low	Low	Low	Low	Low	Low
Nanayakkara et al [28]	Low	Unclear	Low	Low	Low	Low	High	High	Unclear
Meyer et al [29]	Low	Unclear	Low	Low	Low	Low	Low	Low	Unclear
Meiring et al [7]	Low	Low	Low	Low	Low	Low	High	High	Low
Lin et al [30]	Low	Unclear	Low	Low	Low	Low	High	High	Unclear
Krishnan et al [31]	Low	Low	Low	Low	Low	Low	High	High	Low
Kang et al [32]	Low	Unclear	Low	Low	Low	Low	High	High	Unclear
Johnson et al [33]	Low	Low	Low	Low	Low	Low	Low	Low	Low
Holmgren et al [34]	Low	Low	Low	Low	Unclear	Low	High	High	Low
Garcia-Gallo et al [35]	Low	Unclear	Low	Low	Low	Low	High	High	Unclear
El-Rashidy et al [36]	Low	Low	Low	Low	Low	Low	Low	Low	Low
Silva et al [37]	Low	Low	Low	Low	Low	Low	High	High	Low
Caicedo-Torres et al [38]	Low	Low	Low	Low	Low	Low	High	High	Low
Deshmukh et al [39]	Low	Unclear	Low	Low	Low	Low	High	High	Unclear
Ryan et al [40]	Low	Low	Unclear	Low	Low	Low	Low	Unclear	Low
Mayaud et al [41]	Low	Unclear	Low	Unclear	Low	Low	High	High	Unclear

[a]ROB: risk of bias.

[b]Low risk: no relevant shortcomings in ROB assessment.

[c]Unclear risk: unclear ROB in at least one domain and all other domains at low ROB.

[d]High risk: relevant shortcomings in the ROB assessment, at least one domain with high ROB, or model developed without external validation.

Meta-analysis

Forest plots for the NN, Ensemble, SOFA, SAPS II, and APACHE-II models and the associated heterogeneity tests are shown in Figures 3-7. The forest plots and tests of heterogeneity for SVM, NN, DT, and Ensemble models that were not externally validated can be seen in Multimedia Appendix 2. The AUROC for each model type was weighted using the inverse of its variance. Most of the 95% CIs of AUROC estimates from various studies did not overlap within the forest plot; considerable variation among AUROC estimates for both ML and severity of illness score model types was noted. Regrading tests of heterogeneity, I^2 varied between 99% and 100%, τ^2 ranged from 0.0003 to 0.0034, and P was consistently <.01. In Figures 3-7 and Multimedia Appendix 2, the gray boxes represent the weight estimates of the AUROC value from each study. The horizontal line through each gray box illustrates the 95% CI of the AUROC value from that study. Black horizontal lines through a gray box indicate that the CI limits exceeded the length of the gray box. White horizontal lines represent the CI limits that are within the length of the gray box. I^2, τ^2, and Cochran Q P value (denoted as P) are heterogeneity tests.

Random-effects meta-analysis results of the computed pooled AUROC of the ML subgroup models that were externally validated or benchmarked NNs and Ensemble are shown in Figure 3 and Figure 4.

Figure 3. Meta-analysis results: pooled AUROC for externally validated Ensemble models. Gray boxes represent the fixed weight estimates of the AUROC value from each study. Larger gray boxes represent larger fixed weight estimates of the AUROC values. The horizontal line through each gray box illustrates the 95% CI of the AUROC value from that study. Black horizontal lines through a gray box indicate that the CI limits exceed the length of the gray box. White horizontal lines represent CI limits that are within the length of the gray box. The vertical dashed lines in the forest plot are the estimated random pooled effect of the AUROC value from the random-effects meta-analysis. The gray diamonds illustrate the 95% CI for the random pooled effects. Tests of heterogeneity included I^2, τ^2, and Cochran Q *P* value (denoted as P). AUROC: area under the receiver operating curve;.

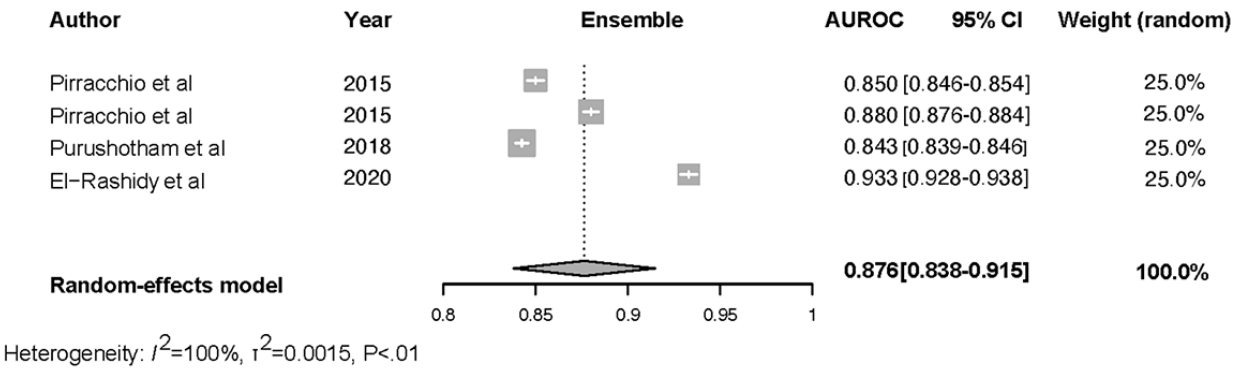


Figure 4. Meta-analysis results: pooled AUROC for externally validated NN models. Gray boxes represent the fixed weight estimates of the AUROC value from each study. Larger gray boxes represent larger fixed weight estimates of the AUROC values. The horizontal line through each gray box illustrates the 95% CI of the AUROC value from that study. Black horizontal lines through a gray box indicate that the CI limits exceed the length of the gray box. White horizontal lines represent CI limits that are within the length of the gray box. The vertical dashed lines in the forest plot are the estimated random pooled effect of the AUROC value from the random-effects meta-analysis. The gray diamonds illustrate the 95% CI for the random pooled effects. Tests of heterogeneity included I^2, τ^2, and Cochran Q *P* value (denoted as P). AUROC: area under the receiver operating curve; NN: neural network.

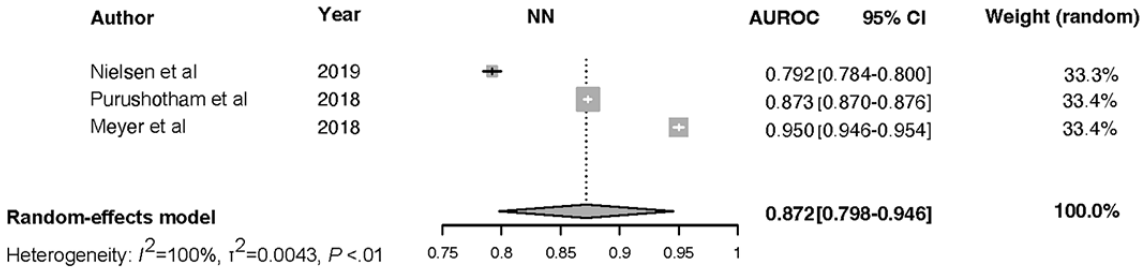


Figure 5. Meta-analysis results: pooled AUROC for SAPS-II. Gray boxes represent the fixed weight estimates of the AUROC value from each study. Larger gray boxes represent larger fixed weight estimates of the AUROC values. The horizontal line through each gray box illustrates the 95% CI of the AUROC value from that study. Black horizontal lines through a gray box indicate that the CI limits exceed the length of the gray box. White horizontal lines represent CI limits that are within the length of the gray box. The vertical dashed lines in the forest plot are the estimated random pooled effect of the AUROC value from the random-effects meta-analysis. The gray diamonds illustrate the 95% CI for the random pooled effects. Tests of heterogeneity included I^2, τ^2, and Cochran Q *P* value (denoted as P). AUROC: area under the receiver operating curve; SAPS-II: Simplified Acute Physiology Score II.

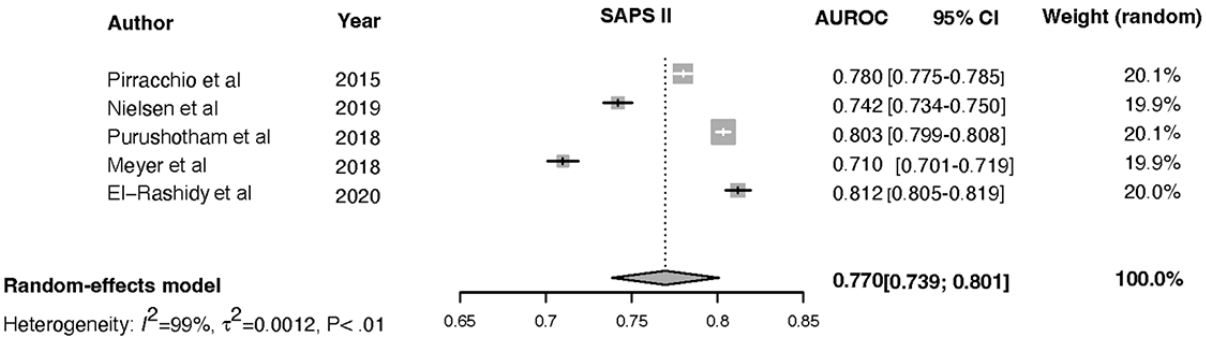


Figure 6. Meta-analysis results: pooled AUROC for SOFA. Gray boxes represent the fixed weight estimates of the AUROC value from each study. Larger gray boxes represent larger fixed weight estimates of the AUROC values. The horizontal line through each gray box illustrates the 95% CI of the AUROC value from that study. Black horizontal lines through a gray box indicate that the CI limits exceed the length of the gray box. White horizontal lines represent CI limits that are within the length of the gray box. The vertical dashed lines in the forest plot are the estimated random pooled effect of the AUROC value from the random-effects meta-analysis. The gray diamonds illustrate the 95% CI for the random pooled effects. Tests of heterogeneity included I^2, τ^2, and Cochran Q *P* value (denoted as P). AUROC: area under the receiver operating curve; SOFA: Sequential Organ Failure Assessment.

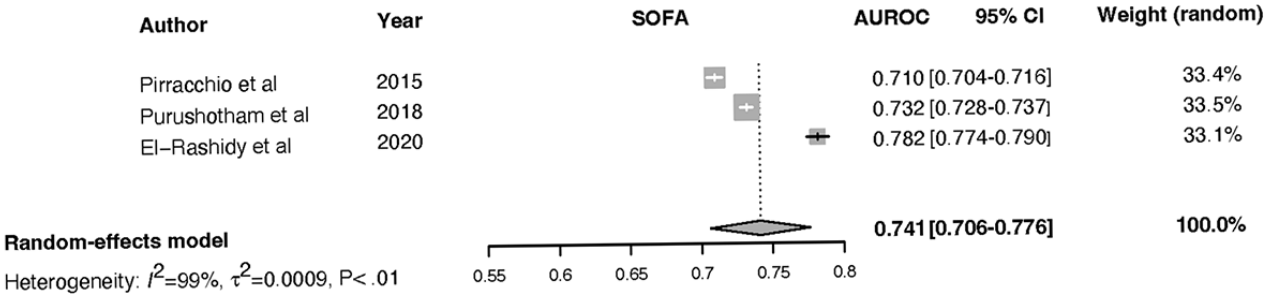
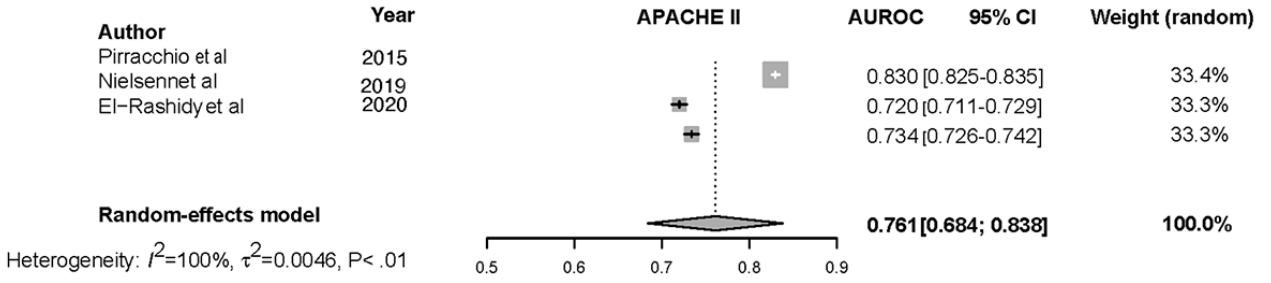


Figure 7. Meta-analysis results: pooled AUROC for APACHE-II. Gray boxes represent the fixed weight estimates of the AUROC value from each study. Larger gray boxes represent larger fixed weight estimates of the AUROC values. The horizontal line through each gray box illustrates the 95% CI of the AUROC value from that study. Black horizontal lines through a gray box indicate that the CI limits exceed the length of the gray box. White horizontal lines represent CI limits that are within the length of the gray box. The vertical dashed lines in the forest plot are the estimated random pooled effect of the AUROC value from the random-effects meta-analysis. The gray diamonds illustrate the 95% CI for the random pooled effects. Tests of heterogeneity included I^2, τ^2, and Cochran Q *P* value (denoted as P). APACHE-II: Acute Physiology and Chronic Health Evaluation-II; AUROC: area under the receiver operating curve;.



The results of heterogeneity for the NN models were as follows: τ^2= 0.0043 (95% CI 0.0014-0.2100), I^2=99.9% (95% CI 99.8%-99.9%), *P*<.01. The results of heterogeneity for the Ensemble models were as follows:

τ^2=0.0015 (95% CI 0.0005-0.0223), I^2=99.7% (95% CI 99.6%-99.8%), *P*<.01. The results were synthesized, and the models are presented in Figure 3 and Figure 4. The results of heterogeneity for the APACHE-2 models were as follows: τ^2=0.0046 (95% CI 0.0011-0.1681), I^2=99.7% (95% CI 99.6%-99.8%), *P*<.01. The results of heterogeneity for the SAPS-II models were as follows: τ^2=0.0012 (95% CI 0.0005-0.0133), I^2=99.2% (95% CI 98.9%-99.4%), *P*<.01. The results of heterogeneity for the SOFA models were as follows: τ^2=0.0009 (95% CI 0.0003-0.0461), I^2=99.1% (95% CI 98.5%-99.4%), *P*<.01 (Figures 5-7).

Discussion

Principal Findings

This is the first study to critically appraise the literature comparing the ML and severity of illness score models to predict ICU mortality. In the reviewed articles, the AUROC of the ML models demonstrated very good discrimination. The range of the ML model AUROC was superior to that of the severity of illness score AUROC. The meta-analysis demonstrated a high degree of heterogeneity and variability within and among studies; therefore, the AUROC performances of the ML and severity of illness score models cannot be pooled, and the results cannot be generalized. Every I^2 value is >97.7%; most of the 95% CIs of AUROC estimates from various studies did not overlap within the forest plot, suggesting considerable variation among AUROC estimates for model types. The CI for AUROC and the statistical significance of the difference in model performance were inconsistently reported within studies. The high heterogeneity came from the diverse study population and practice location, age of inclusion, primary pathology, medical management leading to the ICU admission, and time prediction

window. The heterogenous data management (granularity, frequency of data input, data management, number of predicting variables, prediction timeframe, time series analysis, and training set imbalance) affected model development. It may have resulted in bias, primarily in studies where it has not been addressed (Table 2). Generally, authors reported the ML algorithms with predictive power superior to the clinical scoring system (Table 3); the number of ML models with inferior performance not reported is unknown, which raises the concern of reporting bias. The classification measures of performance were inconsistently reported and required a predefined probability threshold; therefore, models showed different sensitivity and specificity based on the chosen threshold. The variations in the prevalence of the studied outcome secondary to imbalanced data sets make the interpretation of the accuracy difficult. The models' calibration cannot be interpreted because of limited reporting. The external validation process that is necessary to establish generalization was lacking in 65% (13/20) of studies (Table 2). The limited and variable performance metrics reported precludes a comprehensive model performance comparison among studies. The decision curve analysis and model interpretability (explainability) that are necessary to promote transparency and understanding of the model's predictive reasoning was addressed in 25% (5/20) of studies. Results of the clinical performance of ML mortality prediction models as alternatives to the severity of illness score are scarce.

The reviewed studies inconsistently and incompletely captured the descriptive characteristics and other method parameters for ML-based predictive model development. Therefore, we cannot fully assess the superiority or inferiority of ML-based ICU mortality prediction compared with traditional models; however, we recognize the advantage that flexibility in model design offers in the ICU setting.

Study Limitations

This review included studies that were retrospective analyses of data sets with known outcome distributions and incorporated the results of interventions. It is unclear which models were developed exclusively for research purposes; hence, they were not validated. We evaluated studies that compared ML-based mortality prediction models with the severity of illness score–based models, although these models relied on different development statistical methods, variable collection times, and outcome measurement methodologies (SOFA).

The comparison between the artificial intelligence (AI) and severity of illness score models relies only on AUROC values as measures of calibration, discrimination, and classification are not uniformly reported. The random-effects meta-analysis was limited to externally validated models. Owing to the level of heterogeneity, the performance results for most AI and severity of illness score models could not be pooled. The authors recognize that 25% (5/20) of the articles were published between 2004 and 2015 before the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) recommendations for model development and reporting [18]; thus, they were not aligned with the guidelines.

The reviewers assessed the models' ROB and applicability and were aware of the risk of reporting and publication bias favoring the ML models. However, the high heterogeneity among studies prevents an unambiguous interpretation of the funnel plot.

Conclusions and Recommendations

The results of our analysis show that the reporting methodology is incomplete, nonadherent to the current recommendations, and consistent with previous observations [16,50]. The lack of consistent reporting of the measures of the reliability calibration (Brier score and calibration curve of reliability deviation), discrimination, and classification of the probabilistic estimates on external data makes the comparative effectiveness of risk prediction models challenging and has been noted by other authors [43].

Predictive models of mortality can substantially increase patient safety, and by incorporating subtle changes in organ functions that affect outcomes, these models support the early recognition and diagnosis of patients who are deteriorating, thus providing clinicians with additional time to intervene. The heterogeneity of the classification models that was revealed in detail in this review underlines the importance of recognizing the models' ability for temporal and geographical generalization or proper adaptation to previously unseen data [51]. These concepts apply to both models; similar to the ML models, severity of illness score requires periodical updates and customizations to reflect changes in medical care and regional case pathology over time [6].

Our findings lead to the following recommendations for model developers:

1. State whether the developed ML models are intended for clinical practice
2. If models are intended for clinical applications, provide full transparency of the clinical setting from which the data are acquired and all the model development steps; validate the models externally to ensure generalizability
3. If intended for clinical practice, report models' performance metrics, which include measures of discrimination, calibration, and classification, and attach explainer models to facilitate interpretability

Before using ML and/or severity of illness score models as decision support systems to guide clinical practice, we make the following recommendations for clinicians:

1. Be cognizant of the similarities or discrepancies between the cohort used for model development and the local practice population, the practice setting, the model's ability to function prospectively, and the models' lead times
2. Acquire knowledge of the model's performance during testing in the local practice
3. Ensure that the model is periodically updated to changes in patient characteristics and/or clinical variables and adjusted to new clinical practices and therapeutics
4. Confirm that the models' data are monitored and validated and that the model's performance is periodically updated
5. When both the severity of illness score and ML models are available, determine one model's superiority and clinical reliability versus the other through randomized controlled trials

6. When ML models guide clinical practice, ensure that the model makes the correct recommendation for the right reasons and consult the explainer model
7. Identify clinical performance metrics that evaluate the impact of the AI tool on the quality of care, efficiency, productivity, and patient outcomes and account for variability in practice

AI developers must search for and clinicians must be cognizant of the unintended consequences of AI tools; both must understand human–AI tool interactions. Healthcare organization administrators must be aware of the safety, privacy, causality, and ethical challenges when adopting AI tools and recognize the Food and Drug Administration guiding principles for AI/ML development [52].

Acknowledgments

The statistical analysis was funded by the Department of Anesthesiology at Northwestern University.

The authors would like to thank Dr Shaun Grannis for reviewing the manuscript and funding the publication cost.

Conflicts of Interest

None declared.

Multimedia Appendix 1
R scripts for meta-analysis.
[PDF File (Adobe PDF File), 59 KB - medinform_v10i5e35293_app1.pdf]

Multimedia Appendix 2
Forest plots for neural network, decision tree, support vector machine, and Ensemble-based models.
[PDF File (Adobe PDF File), 257 KB - medinform_v10i5e35293_app2.pdf]

References

1. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med 2015 Jan;3(1):42-52 [FREE Full text] [doi: 10.1016/S2213-2600(14)70239-5] [Medline: 25466337]
2. Salluh JI, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. Curr Opin Crit Care 2014 Oct;20(5):557-565. [doi: 10.1097/MCC.0000000000000135] [Medline: 25137401]
3. Keegan MT, Gajic O, Afessa B. Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0III and influence of resuscitation status on model performance. Chest 2012 Oct;142(4):851-858 [FREE Full text] [doi: 10.1378/chest.11-2164] [Medline: 22499827]
4. Breslow MJ, Badawi O. Severity scoring in the critically ill: part 1--interpretation and accuracy of outcome prediction scoring systems. Chest 2012 Jan;141(1):245-252. [doi: 10.1378/chest.11-0330] [Medline: 22215834]
5. Sarkar R, Martin C, Mattie H, Gichoya JW, Stone DJ, Celi LA. Performance of intensive care unit severity scoring systems across different ethnicities. medRxiv (forthcoming) 2021 Jan 20 [FREE Full text] [doi: 10.1101/2021.01.19.21249222] [Medline: 33501459]
6. Strand K, Flaatten H. Severity scoring in the ICU: a review. Acta Anaesthesiol Scand 2008 Apr;52(4):467-478. [doi: 10.1111/j.1399-6576.2008.01586.x] [Medline: 18339152]
7. Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. PLoS One 2018 Nov 14;13(11):e0206862 [FREE Full text] [doi: 10.1371/journal.pone.0206862] [Medline: 30427913]
8. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. Crit Care Med 2011 Jan;39(1):163-169. [doi: 10.1097/CCM.0b013e3181f96f81] [Medline: 20838329]
9. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. Crit Care Med 2018 Jun;46(6):e481-e488. [doi: 10.1097/CCM.0000000000003011] [Medline: 29419557]
10. Balkan B, Essay P, Subbian V. Evaluating ICU clinical severity scoring systems and machine learning applications: APACHE IV/IVa case study. In: Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2018 Presented at: EMBC '18; Honolulu, HI, USA; July 18-21, 2018 p. 4073-4076. [doi: 10.1109/embc.2018.8513324]
11. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review. Crit Care 2008;12(6):R161 [FREE Full text] [doi: 10.1186/cc7160] [Medline: 19091120]

12. de Hond AA, Leeuwenberg AM, Hooft L, Kant IM, Nijman SW, van Os HJ, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022 Jan 10;5(1):2 [FREE Full text] [doi: 10.1038/s41746-021-00549-7] [Medline: 35013569]
13. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814 [FREE Full text] [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]
14. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statist Sci 2001 Aug 1;16(3):199-231. [doi: 10.1214/ss/1009213726]
15. Hassanipour S, Ghaem H, Arab-Zozani M, Seif M, Fararouei M, Abdzadeh E, et al. Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: a systematic review and meta-analysis. Injury 2019 Feb;50(2):244-250. [doi: 10.1016/j.injury.2019.01.007] [Medline: 30660332]
16. Andaur Navarro CL, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281 [FREE Full text] [doi: 10.1136/bmj.n2281] [Medline: 34670780]
17. Bozkurt S, Cahan EM, Seneviratne MG, Sun R, Lossio-Ventura JA, Ioannidis JP, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. J Am Med Inform Assoc 2020 Dec 09;27(12):1878-1884 [FREE Full text] [doi: 10.1093/jamia/ocaa164] [Medline: 32935131]
18. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
19. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes 2020 Oct;13(10):e006556 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.120.006556] [Medline: 33079589]
20. 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. Int J Surg 2021 Apr;88:105906. [doi: 10.1016/j.ijsu.2021.105906] [Medline: 33789826]
21. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017 Jan 05;356:i6460. [doi: 10.1136/bmj.i6460] [Medline: 28057641]
22. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744 [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
23. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 01;170(1):W1-33 [FREE Full text] [doi: 10.7326/M18-1377] [Medline: 30596876]
24. Nielsen AB, Thorsen-Meyer HC, Belling K, Nielsen AP, Thomas CE, Chmura PJ, et al. Survival prediction in intensive-care units based on aggregation of long-term disease history and acute physiology: a retrospective study of the Danish National Patient Registry and electronic patient records. Lancet Digit Health 2019 Jun;1(2):e78-e89 [FREE Full text] [doi: 10.1016/S2589-7500(19)30024-X] [Medline: 33323232]
25. Nimgaonkar A, Karnad DR, Sudarshan S, Ohno-Machado L, Kohane I. Prediction of mortality in an Indian intensive care unit. Comparison between APACHE II and artificial neural networks. Intensive Care Med 2004 Feb;30(2):248-253. [doi: 10.1007/s00134-003-2105-4] [Medline: 14727015]
26. Xia J, Pan S, Zhu M, Cai G, Yan M, Su Q, et al. A long short-term memory ensemble approach for improving the outcome prediction in intensive care unit. Comput Math Methods Med 2019 Nov 3;2019:8152713 [FREE Full text] [doi: 10.1155/2019/8152713] [Medline: 31827589]
27. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. J Biomed Inform 2018 Jul;83:112-134 [FREE Full text] [doi: 10.1016/j.jbi.2018.04.007] [Medline: 29879470]
28. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: a retrospective international registry study. PLoS Med 2018 Nov;15(11):e1002709 [FREE Full text] [doi: 10.1371/journal.pmed.1002709] [Medline: 30500816]
29. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018 Dec;6(12):905-914. [doi: 10.1016/S2213-2600(18)30300-X] [Medline: 30274956]
30. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform 2019 May;125:55-61. [doi: 10.1016/j.ijmedinf.2019.02.002] [Medline: 30914181]
31. Krishnan GS, Kamath SS. A supervised learning approach for ICU mortality prediction based on unstructured electrocardiogram text reports. In: Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems. 2018 May 22 Presented at: NLDB '18; June 13-15, 2018; Paris, France p. 126-134. [doi: 10.1007/978-3-319-91947-8_13]

32. Kang MW, Kim J, Kim DK, Oh K, Joo KW, Kim YS, et al. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. Crit Care 2020 Feb 06;24(1):42 [FREE Full text] [doi: 10.1186/s13054-020-2752-7] [Medline: 32028984]
33. Johnson AE, Kramer AA, Clifford GD. A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. Crit Care Med 2013 Jul;41(7):1711-1718. [doi: 10.1097/CCM.0b013e31828a24fe] [Medline: 23660729]
34. Holmgren G, Andersson P, Jakobsson A, Frigyesi A. Artificial neural networks improve and simplify intensive care mortality prognostication: a national cohort study of 217,289 first-time intensive care unit admissions. J Intensive Care 2019;7:44 [FREE Full text] [doi: 10.1186/s40560-019-0393-1] [Medline: 31428430]
35. García-Gallo JE, Fonseca-Ruiz NJ, Celi LA, Duitama-Muñoz JF. A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis. Med Intensiva (Engl Ed) 2020 Apr;44(3):160-170. [doi: 10.1016/j.medin.2018.07.016] [Medline: 30245121]
36. El-Rashidy N, El-Sappagh S, Abuhmed T, Abdelrazek S, El-Bakry HM. Intensive care unit mortality prediction: an improved patient-specific stacking ensemble model. IEEE Access 2020;8:133541-133564. [doi: 10.1109/access.2020.3010556]
37. Silva A, Cortez P, Santos MF, Gomes L, Neves J. Mortality assessment in intensive care units via adverse events using artificial neural networks. Artif Intell Med 2006 Mar;36(3):223-234. [doi: 10.1016/j.artmed.2005.07.006] [Medline: 16213693]
38. Caicedo-Torres W, Gutierrez J. ISeeU: visually interpretable deep learning for mortality prediction inside the ICU. J Biomed Inform 2019 Oct;98:103269 [FREE Full text] [doi: 10.1016/j.jbi.2019.103269] [Medline: 31430550]
39. Deshmukh F, Merchant SS. Explainable machine learning model for predicting GI bleed mortality in the intensive care unit. Am J Gastroenterol 2020 Oct;115(10):1657-1668. [doi: 10.14309/ajg.0000000000000632] [Medline: 32341266]
40. Ryan L, Lam C, Mataraso S, Allen A, Green-Saxena A, Pellegrini E, et al. Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: a retrospective study. Ann Med Surg (Lond) 2020 Nov;59:207-216 [FREE Full text] [doi: 10.1016/j.amsu.2020.09.044] [Medline: 33042536]
41. Mayaud L, Lai PS, Clifford GD, Tarassenko L, Celi LA, Annane D. Dynamic data during hypotensive episode improves mortality predictions among patients with sepsis and hypotension. Crit Care Med 2013 Apr;41(4):954-962 [FREE Full text] [doi: 10.1097/CCM.0b013e3182772adb] [Medline: 23385106]
42. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]
43. Huang C, Li S, Caraballo C, Masoudi FA, Rumsfeld JS, Spertus JA, et al. Performance metrics for the comparative analysis of clinical risk prediction models employing machine learning. Circ Cardiovasc Qual Outcomes 2021 Oct;14(10):e007526. [doi: 10.1161/CIRCOUTCOMES.120.007526] [Medline: 34601947]
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. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: a systematic review and meta-analysis. Int J Med Inform 2021 Jul;151:104484 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104484] [Medline: 33991886]
46. 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]
47. Louie KS, Seigneurin A, Cathcart P, Sasieni P. Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis. Ann Oncol 2015 May;26(5):848-864 [FREE Full text] [doi: 10.1093/annonc/mdu525] [Medline: 25403590]
48. Hosmer DW, Lemeshow S. Assessing the fit of the model: 5.2.4. area under the receiver operating characteristic curve. In: Shewhart WA, Wilks SS, editors. Applied Logistic Regression. 2nd Edition. Hoboken, NJ, US: John Wiley & Sons; Sep 13, 2000:143-202.
49. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Review of Interventions. London, UK: The Cochrane Collaboration; 2021.
50. Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med Inform Decis Mak 2021 Feb 15;21(1):54 [FREE Full text] [doi: 10.1186/s12911-021-01403-2] [Medline: 33588830]
51. Futoma J, Simons M, Doshi-Velez F, Kamaleswaran R. Generalization in clinical prediction models: the blessing and curse of measurement indicator variables. Crit Care Explor 2021 Jul;3(7):e0453 [FREE Full text] [doi: 10.1097/CCE.0000000000000453] [Medline: 34235453]
52. Good Machine Learning Practice for Medical Device Development: Guiding Principles. U.S. Food & Drug Administration. 2021. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles [accessed 2022-05-08]

Abbreviations

AI: artificial intelligence
APACHE: Acute Physiology and Chronic Health Evaluation
AUROC: area under the receiver operating curve
CHARMS: Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies
ICU: intensive care unit
ML: machine learning
NN: neural network
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST: Prediction model Risk of Bias Assessment Tool
ROB: risk of bias
SAPS: Simplified Acute Physiology Score
SOFA: Sequential Organ Failure Assessment
SVM: support vector machine
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 29.11.21; peer-reviewed by P Giabbanelli, R Gore; comments to author 13.02.22; revised version received 24.04.22; accepted 25.04.22; published 31.05.22.

Please cite as:
Barboi C, Tzavelis A, Muhammad LN
Comparison of Severity of Illness Scores and Artificial Intelligence Models That Are Predictive of Intensive Care Unit Mortality: Meta-analysis and Review of the Literature
JMIR Med Inform 2022;10(5):e35293
URL: https://medinform.jmir.org/2022/5/e35293
doi:10.2196/35293
PMID:35639445

Review

Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review

Jonathan Huang[1], BSc; Galal Galal[1], MD, MPH; Mozziyar Etemadi[1,2], MD, PhD; Mahesh Vaidyanathan[1,3], MD, MBA

[1]Department of Anesthesiology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[2]Department of Biomedical Engineering, Northwestern University, Evanston, IL, United States

[3]Digital Health & Data Science Curricular Thread, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

Corresponding Author:
Galal Galal, MD, MPH
Department of Anesthesiology
Northwestern University Feinberg School of Medicine
420 E Superior St
Chicago, IL, 60611
United States
Phone: 1 (312) 503 8194
Email: galal.galal@nm.org

Abstract

Background: Racial bias is a key concern regarding the development, validation, and implementation of machine learning (ML) models in clinical settings. Despite the potential of bias to propagate health disparities, racial bias in clinical ML has yet to be thoroughly examined and best practices for bias mitigation remain unclear.

Objective: Our objective was to perform a scoping review to characterize the methods by which the racial bias of ML has been assessed and describe strategies that may be used to enhance algorithmic fairness in clinical ML.

Methods: A scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Extension for Scoping Reviews. A literature search using PubMed, Scopus, and Embase databases, as well as Google Scholar, identified 635 records, of which 12 studies were included.

Results: Applications of ML were varied and involved diagnosis, outcome prediction, and clinical score prediction performed on data sets including images, diagnostic studies, clinical text, and clinical variables. Of the 12 studies, 1 (8%) described a model in routine clinical use, 2 (17%) examined prospectively validated clinical models, and the remaining 9 (75%) described internally validated models. In addition, 8 (67%) studies concluded that racial bias was present, 2 (17%) concluded that it was not, and 2 (17%) assessed the implementation of bias mitigation strategies without comparison to a baseline model. Fairness metrics used to assess algorithmic racial bias were inconsistent. The most commonly observed metrics were equal opportunity difference (5/12, 42%), accuracy (4/12, 25%), and disparate impact (2/12, 17%). All 8 (67%) studies that implemented methods for mitigation of racial bias successfully increased fairness, as measured by the authors' chosen metrics. Preprocessing methods of bias mitigation were most commonly used across all studies that implemented them.

Conclusions: The broad scope of medical ML applications and potential patient harms demand an increased emphasis on evaluation and mitigation of racial bias in clinical ML. However, the adoption of algorithmic fairness principles in medicine remains inconsistent and is limited by poor data availability and ML model reporting. We recommend that researchers and journal editors emphasize standardized reporting and data availability in medical ML studies to improve transparency and facilitate evaluation for racial bias.

(JMIR Med Inform 2022;10(5):e36388) doi:10.2196/36388

KEYWORDS

artificial intelligence; machine learning; race; bias; racial bias; scoping review; algorithm; algorithmic fairness; clinical machine learning; medical machine learning; fairness; assessment; model; diagnosis; outcome prediction; score prediction; prediction; mitigation

Introduction

Background

In recent years, artificial intelligence (AI) has drawn significant attention in medicine as machine learning (ML) techniques show an increasing promise of clinical impact. Driven by unprecedented data accessibility and computational capacity, ML has been reported to reach parity with human clinicians in a variety of tasks [1-3]. ML is poised to benefit patients and physicians by optimizing clinical workflows, enhancing diagnosis, and supporting personalized health care interventions [4-6]. Decision support tools based on ML have already been implemented across health systems [7,8], and the continued proliferation of clinical ML will impact patients in all fields of medicine.

However, despite its appeal, significant barriers remain to the full realization of clinically integrated ML. Key concerns include limited model transparency due to the "black box" of ML, inadequate reporting standards, and the need for prospective validation in clinical settings [1,9-12]. Racial bias in clinical ML is a crucial challenge arising from these limitations and must be addressed to ensure fairness in clinical implementation of ML. As ML is premised on prediction of novel outcomes based on previously seen examples, unintended discrimination is a natural consequence of algorithm development involving training data that reflect real-world inequities [13].

Equity in health care remains a continual pursuit [14,15]. Bias and disparities along dimensions of race, age, and gender have been shown to impact health care access and delivery, evident in varied settings, such as race correction in clinical algorithms or clinical trial enrollment and adverse event monitoring [16,17]. Considering the growing body of literature demonstrating profound adverse impacts of health care inequities on patient outcomes, mitigation of the numerous and insidious sources of potential bias in medicine requires remains a critical challenge to prevent harm to patients [14,17]. Thus, the potential for algorithms to perpetuate health disparities must be carefully weighed when incorporating ML models into clinical practice [18-20].

Algorithmic fairness is an area of ML research guiding model development with the aim of preventing discrimination involving protected groups, which are defined by attributes such as race, gender, religion, physiologic variability, preexisting conditions, physical ability, and sexual orientation [13,19]. However, application of algorithmic fairness principles in the medical ML literature remains nascent [20]. Greater awareness of the potential harms of bias in clinical ML as well as methods to evaluate and mitigate them is needed to support clinicians and researchers across the health care and data science disciplines, who must evaluate and implement clinical ML models with a critical eye toward algorithmic fairness. The objective of this study is to characterize the impact and mitigation of racial bias in clinical ML to date and describe best practices for research efforts extending algorithmic fairness to medicine.

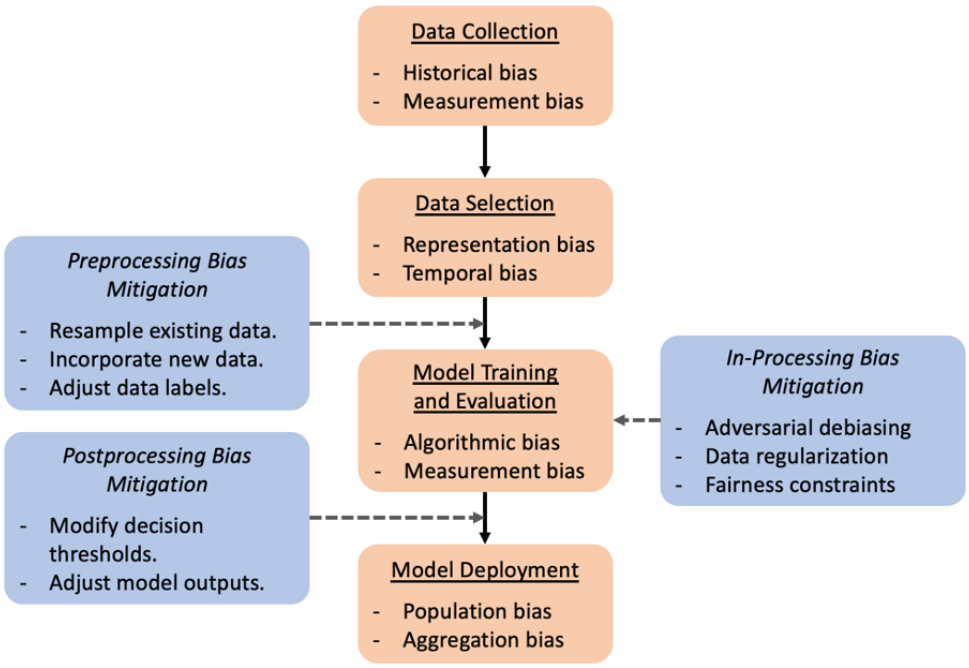
Bias and Fairness in Machine Learning

In the setting of algorithmic fairness, bias is present when an algorithm systematically favors one outcome over another. Bias may be introduced into an ML algorithm throughout all steps of the development process, which involves data collection, data selection, model training, and model deployment [13]. Examples of these sources of bias are shown in Figure 1, and their definitions are given in Multimedia Appendix 1. Notably, historical bias may be present even if all steps of model development are optimally performed. This is of particular concern in the evaluation of racial bias in clinical ML, given the presence of existing and historical health care disparities [14].

Depending on the context, bias in clinical ML may not be harmful and can even be used to overcome inequality [13]. In situations in which targeting a well-defined subpopulation above all others is desirable, an ML algorithm biased toward a particular group may be used to proactively mitigate existing disparities. However, bias may arise when ML models designed to serve the needs of a specific clinical population—such as a particular community or high-risk demographic—are inappropriately applied to other populations or when more general models are applied to specific populations. Additionally, ML algorithms tend to overfit to the data on which they are trained, which entails the learning of spurious relationships present in the training data set and may result in a lack of generalizability to other settings. As a result, a model that appears unbiased in one setting may display bias in another. Thus, bias in clinical ML must be considered in the light of the context and particular population of interest.

Bias in an ML model may lead to unfairness if not appropriately evaluated and accounted for. Fairness in ML is achieved when algorithmic decision-making does not favor an individual or group based on protected attributes. Research efforts have emphasized group fairness over individual fairness, given the need for algorithms that consider existing differences between populations—whether intrinsic or extrinsic—while preventing discrimination between groups [13,21]. Crucially, improving model fairness does not necessarily require compromising accuracy overall [22]. For instance, an unfair disease-screening tool might have poor sensitivity for disease detection in one low-risk population subgroup compared to another with higher risk; improving the fairness of this tool would entail adjusting the model to have more similar sensitivities between subgroups. In this study, we examine the racial bias of clinical ML in terms of model fairness with respect to race.

Figure 1. The clinical machine learning development workflow (orange boxes) offers several opportunities (blue boxes) to evaluate and mitigate potential biases introduced by the data set or model. Preprocessing methods seek to adjust the existing data set to preempt biases resulting from inadequate data representation or labeling. In-processing methods impose fairness constraints as additional metrics optimized by the model during training or present data in a structured manner to avoid biases in the sampling process. Postprocessing methods account for model biases by adjusting model outputs or changing the way they are used.



Assessing and Achieving Fairness in Machine Learning

Group fairness is quantified by evaluating the similarity of a given statistical metric between predictions made for different groups. Group fairness indicators encountered in this review are defined in Table 1. Critical examinations of different methods for evaluating fairness in ML, both in general application [13,23,24] and in the context of health care [21], have been previously described, though applications in clinical ML remain limited. It is important to note that fairness metrics may be at odds with one another, depending on the context and application [25]; thus, evaluation of an appropriate metric, given the clinical situation of interest, is paramount [26].

Approaches to bias mitigation fall into 3 major categories (Figure 1): *preprocessing*, in which inequities in data are removed prior to model training; *in-processing*, in which the model training process is conducted to actively prevent discrimination; and *postprocessing*, in which outputs of a trained model are adjusted to achieve fairness [13]. Preprocessing can be performed by resampling existing data, incorporating new data, or adjusting data labels. In-processing methods use adversarial techniques, impose constraints and regularization, or ensure fairness of underlying representations during training. Finally, postprocessing entails group-specific modification of decision thresholds or outcomes to ensure fairness in the application of model predictions. Different approaches may be optimal depending on the setting and stage of model development.

Table 1. Group fairness metrics encountered in this review.

Term	Description
AUROC[a]	Assesses overall classifier performance by measuring the TPR[b] and FPR[c] of a classifier at different thresholds.
Average odds	Compares the average of the TPR and FPR for the classification outcome between protected and unprotected groups.
Balanced accuracy	A measure of accuracy corrected for data imbalance, calculated as the average of sensitivity and specificity for a group.
Calibration	Assesses how well the risk score or probability predictions reflect actual outcomes.
Disparate impact	Measures deviation from statistical parity, calculated as the ratio of the rate of the positive outcome between protected and unprotected groups. Ideally, the disparate impact is 1.
Equal opportunity	For classification tasks in which one outcome is preferred over the other, equal opportunity is satisfied when the preferred outcome is predicted with equal accuracy between protected and unprotected groups. Ideally, the TPR or FNR[d] disparity between groups is 0.
Equalized odds	The TPR and FPR are equal between protected and unprotected groups.
Error rate	Compares the error rate of predictions, calculated as the number of incorrect predictions divided by the total number of predictions, between protected and unprotected groups. Ideally, the error rate disparity between groups is 0.
Statistical parity	Statistical parity (also known as demographic parity) is satisfied when the rate of positive outcomes is equal between protected and unprotected groups.

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

[b]TPR: true-positive rate.

[c]FPR: false-positive rate.

[d]FNR: false-negative rate.

Methods

Study Design

We performed a scoping review of racial bias and algorithmic fairness in clinical ML models in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) 2020 guidelines [27] and PRISMA Extension for Scoping Reviews [28]. The review protocol was not registered and is available upon request to the authors. The PubMed MEDLINE (National Library of Medicine), Scopus (Elsevier), and Embase (Elsevier) databases were queried by combining terminology pertaining to ML, race, and bias as keywords. Additional records were identified using Google Scholar search. The exact search strategy is detailed in Multimedia Appendix 1.

Study Selection

After duplicate record removal, studies were initially screened by title and abstract and then screened for final inclusion by full text review. All screening was performed independently by 2 reviewers. Studies were selected based on the following inclusion criteria: peer-reviewed original research, English language, full text available, development or evaluation of a clinically relevant ML model, and evaluation of bias of the model regarding racial or ethnic groups. Studies other than full-length papers were excluded. ML was defined as a computer algorithm that improves automatically via training on data [4]. Per PRISMA guidelines, any disagreements regarding study inclusion based on these criteria were reconciled by discussion.

Data Abstraction

Relevant data were abstracted from included papers by 1 reviewer. Data of interest included the clinical objective of ML models, identification of racial bias, impact of racial bias, metrics for bias assessment, mitigation of racial bias, methods for bias mitigation, data set size, data source, ML model architecture, and availability of computer code used for data preparation and ML model development. The methodological quality of included studies was not assessed, given the scoping nature of this review [28].

Results

Study Characteristics

The literature search was performed on September 8, 2021, and identified 635 records (Figure 2). Of these, 26 (4.1%) full-text papers were reviewed and 12 (46.2%) were included in the final analysis [29-40].

Characteristics of the included studies are summarized in Table 2. Data sets and models used are summarized in Multimedia Appendix 1. Of the 12 studies, 3 (25%) were published in 2019, 5 (42%) in 2020, and 4 (33%) in 2021. In addition, 9 (75%) studies originated from the United States, 1 (8%) from Canada, 1 (8%) from Sweden, and 1 (8%) from both the United Kingdom and Nigeria. Applications of ML were varied and involved diagnosis, outcome prediction, and clinical score prediction performed on data sets including images, diagnostic studies, clinical text, and clinical variables. Furthermore, 1 (8%) study described a model in routine clinical use [36], 2 (17%) examined prospectively validated clinical models [35,39], and the remaining 9 (75%) described internally validated models.

Of the 12 studies, 5 (42%) published code used for analysis, 3 (25%) made model development code available [34,36,39], 2 (17%) published bias analysis code [33,36], 1 (8%) published code relevant to debiasing [30], and 1 (8%) published data selection code [33]. In addition, 1 (8%) study used publicly available code for analysis [31], and code was specified as

available upon request in 1 (8%) study [35]. Bias of an ML model was evaluated using an external database in 8 (67%) studies [30-34,37,38], single-institutional data in 3 (25%) studies [35,36,40], and data from 2 institutions in 2 (17%) studies [29,39]. No institutional data sets were published. Convolutional neural networks (CNNs) were the predominant ML modeling technique used (5/12, 42%), followed by logistic regression (3/12, 25%), least absolute shrinkage and selection operator (LASSO; 2/12, 17%), and extreme gradient boosting (XGBoost; 2/12, 17%). In addition, 3 (25%) studies evaluated models adapted from existing neural network architectures: ResNet50 in 2 (17%) studies [29,32] and DenseNet in the other [38].

Of the 12 studies, 9 (75%) evaluated a model developed internally by the same researchers [29-33,35,37,39,40], 2 (17%) evaluated a model developed externally by separate researchers [36,38], and 1 (8%) evaluated both internally and externally developed models [34]. In addition, 8 (67%) studies concluded that racial bias was present [29,32-34,36-39], 2 (17%) concluded that bias was not present [35,40], and 2 (17%) assessed the implementation of bias mitigation strategies without comparison to a baseline model [30,31]. A variety of methods were used to assess the presence of algorithmic racial bias: 3 (25%) studies used multiple metrics to assess fairness [31,34,37], while the remaining 9 (75%) used a single metric. The most commonly used fairness metrics were equal opportunity difference [41], defined either as the difference in the true-positive rate (TPR) or the false-negative rate (FNR) between subgroups (5/12, 42%) [30,31,38,39]; accuracy (4/12, 25%) [29,31,32,34]; and disparate impact (2/12, 17%) [31,37].

The approaches and efficacy of bias mitigation methods used in the studies evaluated are summarized in Table 3. All 8 (67%) studies that implemented methods for mitigation of racial bias successfully increased fairness, as measured by the authors' chosen metrics [29-32,34,36,37,39]. Preprocessing bias mitigation was the most commonly used strategy (7/13, 54%). In addition, 1 (8%) study removed race information from the training data, though superior improvements in disparate impact and equal opportunity difference were achieved by reweighing [37]. Furthermore, 2 (17%) studies performed in-processing bias mitigation using the prejudice remover regularizer [42] or adversarial debiasing during model training [31,37]. However, in both studies, in-processing was ineffective in reducing bias and was outperformed by other bias mitigation methods. Finally, 1 (8%) study evaluated multiple types of ML models for bias during the development process, concluding that a LASSO model was preferable to conditional random forest, gradient boosting, and ensemble models for racially unbiased dementia ascertainment [34].

Figure 2. PRISMA flowchart of study inclusion. ML: machine learning; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses.

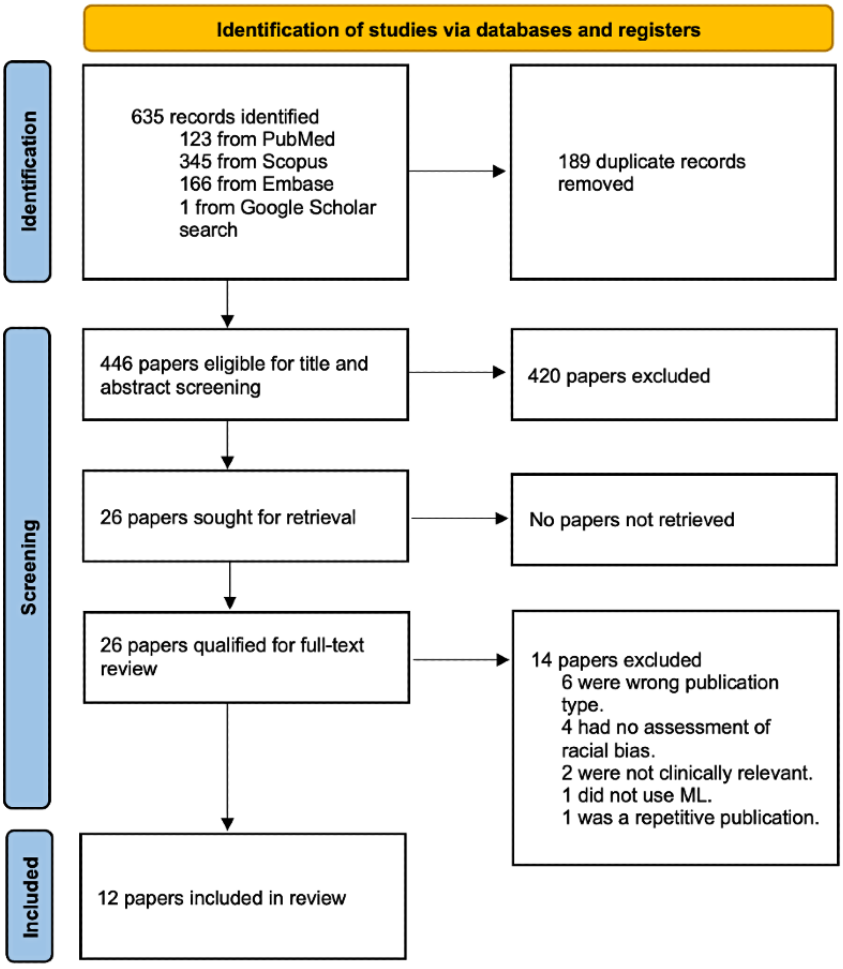


Table 2. Study characteristics.

Author (year)	Clinical objective	How was fairness evaluated?	Was racial bias identified?	How was the AI[a] model biased?	Was racial bias mitigated?	Protected class
Abubakar et al (2020) [29]	Identification of images of burns vs healthy skin	Accuracy	Yes	Poor accuracy of models trained on a Caucasian data set and validated on an African data set and vice versa	Yes	Dark-skinned patients, light-skinned patients
Allen et al (2020) [30]	Intensive care unit (ICU) mortality prediction	Equal opportunity difference (FNR[b] disparity)	N/A[c]	N/A	Yes	Non-White patients
Briggs and Hollmén (2020) [31]	Prediction of future health care expenditures of individual patients	Balanced accuracy, statistical parity, disparate impact, average odds, equal opportunity	N/A	N/A	Yes	Black patients
Burlina et al (2021) [32]	Diagnosis of diabetic retinopathy from fundus photography	Accuracy	Yes	Lower diagnostic accuracy in darker-skinned individuals compared to lighter-skinned individuals	Yes	Dark-skinned patients
Chen et al (2019) [33]	ICU mortality prediction, psychiatric readmission prediction	Error rate (0-1 loss)	Yes	Differences in error rates in ICU mortality between racial groups	No	Non-White patients
Gianattasio et al (2020) [34]	Dementia status classification	Sensitivity, specificity, accuracy	Yes	Existing algorithms varying in sensitivity and specificity between race/ethnicity groups	Yes	Hispanic, non-Hispanic Black patients
Noseworthy et al (2020) [35]	Prediction of left ventricular ejection fraction ≤35% from the electrocardiogram (ECG)	AUROC[d]	No	N/A	No	Non-White patients
Obermeyer et al (2019) [36]	Prediction of future health care expenditures of individual patients	Calibration	Yes	Black patients with a higher burden than White patients at the same algorithmic risk score	Yes	Black patients
Park et al (2021) [37]	Prediction of postpartum depression and postpartum mental health service utilization	Disparate impact, equal opportunity difference (TPR[e] disparity)	Yes	Black women with a worse health status than White women at the same predicted risk level	Yes	Black patients
Seyyed-Kalantari et al (2021) [38]	Diagnostic label prediction from chest X-rays	Equal opportunity difference (TPR disparity)	Yes	Greater TPR disparity in Hispanic patients	No	Non-White patients
Thompson et al (2021) [39]	Identification of opioid misuse from clinical notes	Equal opportunity difference (FNR disparity)	Yes	Greater FNR in the Black subgroup than in the White subgroup	Yes	Black patients
Wissel et al (2019) [40]	Assignment of surgical candidacy score for patients with epilepsy using clinical notes	Regression analysis of the impact of the race variable on the candidacy score	No	N/A	No	Non-White patients

[a]AI: artificial intelligence.

[b]FNR: false-negative rate.

[c]N/A: not applicable.

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

[e]TPR: true-positive rate.

Table 3. Bias mitigation methods among reviewed studies.

Description of strategies used	Effectiveness
Preprocessing	
Reweighing training data	• An equal opportunity difference (FNR[a] difference) of 0.016 (P=.20) was achieved for intensive care unit (ICU) mortality prediction [33]. • The mean fairness measure (average of statistical parity difference, disparate impact measure, average odds difference, and equal opportunity difference) improved to 0.06 from 0.12 for prediction of health care costs [34]. • Disparate impact improved from 0.31 to 0.79, and the equal opportunity (TPR[b]) difference improved from –0.19 to 0.02 for prediction of postpartum depression development; prediction of mental health service use in pregnant individuals improved from 0.45 to 0.85 and –0.11 to –0.02, respectively [40].
Combining data sets to increase heterogeneity	• The accuracy of skin burn identification increased to 99.5% using a combined data set compared to 83.4% and 87.5% when trained on an African and evaluated on a Caucasian data set and vice versa [32].
Generating synthetic minority class data	• Disparity in diabetic retinopathy diagnostic accuracy improved from 12.5% to 7.5% and 0.5% when augmenting with retina appearance-optimized images and diabetic retinopathy status-optimized images created with a generative adversarial network, respectively [35].
Adjusting label selection	• Improved congruence in health outcomes between groups after developing models to predict other labels for health status besides financial expenditures [39].
Removing race information from training data	• Disparate impact improved from 0.31 to 0.61 and equal opportunity (TPR) difference improved from –0.19 to –0.05 for prediction of postpartum depression development; respective improvements from 0.45 to 0.63 and –0.11 to –0.04 for prediction of mental health service use in pregnant individuals [40].
In-processing	
Use of a regularizer during training	• Disparate impact improved, but accuracy and the equal opportunity (TPR) difference decreased when implementing the prejudice remover regularizer in prediction of postpartum depression in pregnant individuals [40].
Adversarial debiasing	• The mean fairness measure (average of statistical parity difference, disparate impact measure, average odds difference, and equal opportunity difference) worsened to 0.07 from 0.05 for prediction of health care costs [34].
Postprocessing	
Calibration	• The equal opportunity (FNR) difference improved from 0.15 to 0.03 for identification of opioid misuse [42].
Reject option-based classification	• The mean fairness measure (average of statistical parity difference, disparate impact measure, average odds difference, and equal opportunity difference) improved to 0.09 from 0.15 for prediction of health care costs [34].
Varying cut-point selection	• The equal opportunity (FNR) difference improved from 0.15 to 0.04 for identification of opioid misuse [42]. • The congruence in sensitivity and specificity between groups improved without reduction in accuracy for classification of dementia status [37].

[a]FNR: false-negative rate.

[b]TPR: true-positive rate.

Discussion

Principal Findings

Given the pressing issue of equity in health care and the rapid development of medical ML applications, racial bias must be thoroughly evaluated in clinical ML models in order to protect patient safety and prevent the algorithmic encoding of inequality. Algorithmic fairness is a relatively novel field within the discipline of ML, and its application to medical ML remains nascent. In our evaluation of the literature describing mitigation of racial bias in clinical ML, we identified a variety of bias mitigation methods, which when applied successfully increase fairness and demonstrate the feasibility and importance of racial bias evaluation in the medical ML development process. Based on our findings, there is a need for heightened awareness of algorithmic fairness concepts, increased data availability, and improved reporting transparency in medical ML development to ensure fairness in clinical ML.

Impact of Racial Bias in Clinical Machine Learning

The broad scope of medical ML applications and potential patient harms following deployment across health care systems demand an increased emphasis on evaluation and mitigation of racial bias in clinical ML. Screening and outcome prediction tasks are commonly examined among reviewed studies. Racial bias in such tasks is particularly concerning as decisions made from flawed models trained on data, which reflect historical inequities in disease diagnosis and care delivery, may perpetuate inequalities by shaping clinical decision-making [14,19]. Evaluation and mitigation of potential biases must occur throughout the model development life cycle to protect patients from algorithmic unfairness.

Reviewed studies frequently identified racial bias in clinical ML models. Notably, 1 algorithm in clinical use for prediction of future health care expenditures was found to discriminate against Black patients when compared to White patients, potentially contributing to disparities in health care delivery [36]. Other ML models that possibly demonstrate racial bias remain in preclinical states of development. Several studies have explicitly studied racial bias against Black patients compared to White patients. For example, 2 studies demonstrated that ML algorithms predicted similar risk scores in Black and White patients, though the Black patients were less healthy [36,37], and another demonstrated that an opioid misuse classifier had a higher FNR for Black patients [39]. Disparities in mortality prediction and X-ray diagnosis were identified in other races and ethnic groups [33,34,38], as well as disparities in burn identification and diabetic retinopathy identification in dark-skinned versus lighter-skinned patients [29,32]. Although conclusions cannot be drawn regarding the prevalence of racial bias among published clinical ML studies, the broad scope of clinical ML models susceptible to racial bias in this review exposes the potential of racial bias encoded in ML models to negatively impact patients across all aspects of health care.

Assessment of Racial Bias

Clinical ML models must be carefully evaluated for potential biases imposed upon patients. Different fairness metrics may highlight different aspects of fairness relevant to a particular clinical setting; therefore, evaluation of all appropriate fairness metrics is needed when evaluating for potential bias. For example, calibration is particularly important to models performing risk prediction, while equal opportunity and disparate impact are relevant to screening and diagnostic settings. Inconsistent choice of fairness metrics among studies included in this review shows the need for a more standardized assessment process of racial bias in clinical ML. Some studies assessed fairness using metrics such as accuracy, area under the receiver operating characteristic curve (AUROC), and correlation of outcome with race, which may not sufficiently evaluate fairness [21]. Moreover, there are inherent trade-offs to the use of different fairness metrics [25], and static fairness criteria may even lead to delayed harms in the long term [43].

Obermeyer et al [36] present an example of using model calibration in conjunction with varied outcome labels to successfully de-bias an algorithm used to manage population health, and case studies have examined trade-offs of bias evaluation metrics in other settings, such as criminal justice [44], which may also serve as useful frameworks for clinical ML researchers. Use of "causal models," which allow for closely tailored examination of discriminatory relationships in data, is another opportunity for investigation and mitigation of biased model behavior [45]. An increased focus from medical journals on bias evaluation checklists applicable to clinical ML models, such as the Prediction Model Risk of Bias Assessment Tool (PROBAST), is desirable to further emphasize vigilance regarding biased ML models [46]. Ultimately, more thorough analysis of fairness criteria in clinical ML will allow researchers to better contextualize and act on potential biases.

Clinical ML researchers should also be aware of potential barriers to ML fairness when adapting pretrained models and data representations. For instance, deep neural networks performing image processing tasks are frequently pretrained on large data sets and then fine-tuned to adapt to other tasks. Methods for removal of spurious variations from such models have been described, such as joint learning and unlearning algorithms, which account for contributions of undesirable variations during model development [47]. Language models trained in an unsupervised manner on vast amounts of text may learn biases present in training data [48]. Similarly, biases have been described in word embeddings [49], which are vectorized word representations used as inputs to ML models. Identification of bias in embeddings raises concerns about performance disparities in clinical applications of natural language processing if the bias is not screened for and appropriately addressed [50]. The lack of interpretability often inherent to ML models heightens the need for thorough evaluation of their potential biases.

Creating Fair Models

Preprocessing and postprocessing methods of bias mitigation were successfully implemented among the publications reviewed for this study. Postprocessing methods appear to be easier to implement and may allow tailoring of imperfect models to new settings [51]. However, using preprocessing and in-processing to create unbiased data sets and algorithms at the outset of model development is desirable to facilitate the creation of fair, generalizable models. Continued evaluation of these techniques in clinical contexts is needed to inform best practices.

As data quality is generally the limiting factor to development of robust ML models, improvements to data generally translates directly into model performance improvements. Supplementation of data sets using generative models to synthesize patient data may be a viable approach to address data limitations. A study by Burlina et al [32] illustrated this fact by using a generative adversarial network to synthesize fundoscopy images while reducing class imbalance. However, though data limitations may contribute to disparities in model performance across racial groups, algorithmic unfairness may arise from other underlying biases in data as well [38]. Publications included in this review demonstrated improved fairness in ML models using multisource data sets, which may mitigate biases in the data collection process of single-source data sets [29,38]. Moreover, care must also be taken to ensure that

multi-institutional data sets are appropriately prepared and used due to evidence that site-specific signatures contribute to bias in ML models [52]. Finally, protected attributes should not simply be ignored during model development, an approach called "fairness through unawareness," as models may be able to infer protected group membership from other data features. Additionally, omission of protected attributes may cause bias if a legitimate relationship exists between the attribute and outcome of interest [19].

Several online resources aggregate examples and code implementations of published fairness evaluation and bias mitigation methods. Some examples of these resources include Aequitas, Artificial Intelligence Fairness 360 (IBM, Armonk, NY, United States), and Fairlearn (Microsoft Corporation, Redmond, WA, United States) [53,54]. Additionally, TensorFlow, a popular deep learning framework, includes a tool for evaluation of fairness indicators. Work by Briggs et al [31] highlights the feasibility and positive impact of standardized methodologies for addressing bias using a variety of performance indicators and mitigation techniques. Greater adoption of these and other strategies in fairness evaluation and bias mitigation will help set standard benchmarks for fairness in clinical ML.

The Role of Transparency and Data Availability

ML is often characterized as a black box due to its limited interpretability, which is particularly problematic when attempting to address and prevent racial biases in clinical ML [55]. Although research in recent years has yielded significant progress in explainable ML methods [56], publication of model development code and data sets remains the most straightforward approach to transparency. Regrettably, medical ML research falls far short of these standards [57,58]. Code and data availability was inconsistent among the publications included in this review, and the majority of studies evaluated racial bias using publicly available data sets, including the Medical Information Mart for Intensive Care (MIMIC) [30,33,38], Kaggle EyePACS [32], and Dissecting Bias [31]. Considering the vast number of private, institutional data sets used to develop clinical ML models, there is a crucial need for future publications to maximize transparency, ensuring the ability to evaluate for fairness in clinical ML.

Increased publication of institutional data sets would facilitate the interdisciplinary collaboration needed to translate concepts of fairness in ML into the realm of medicine. Improved availability of data sets would also enable researchers to more easily validate existing models and perform fairness evaluations on different patient populations, translating benefits of ML across populations. Additionally, collaboration between institutions to maintain diverse, broadly representative data sets would facilitate the development of generalizable models free of the biases inherent to single-institutional data. However, ethical and patient confidentiality considerations may limit publication of clinical data. In contrast, publication of code and trained models, which are infrequently made available in the clinical ML literature [1,59], would similarly allow researchers to assess clinical ML on diverse populations without limitations imposed by patient privacy standards or institutional data-sharing regulations. Another possible paradigm to mitigate bias by training on diversely representative data sets while maintaining data privacy is federated learning, which involves piecewise training of an ML model on separate data sets and removes the need for data sharing during model development [60].

Moreover, increased emphasis on fairness in clinical ML through adoption of model development and reporting guidelines is needed [59,61]. Reporting guidelines for medical ML studies are inconsistently adopted, due in part to a lack of editorial policies among medical journals [1]. Moreover, reporting of demographic information needed to assess biases due to data sets is lacking [62,63]. The proposed Minimum Information for Medical AI Reporting guideline addresses these concerns by recommending that clinical ML studies report information necessary for understanding potential biases, including relevant demographic information of patient data used for model development [64]. In conjunction with upcoming reporting guidelines tailored to clinical ML [61], efforts to improve reporting quality will contribute to a standardized framework for fairness evaluation and bias mitigation in clinical ML.

Limitations

As with any literature review, there are limitations to this study. Given the heterogeneity of terminology used to describe ML and racial bias, our search may have overlooked relevant publications. Additionally, we were limited by publication bias as we excluded publications other than full-length manuscripts, and researchers may be less likely to publish results confirming the absence of racial bias in a clinical ML model. Finally, the novelty of ML fairness in medicine and the resulting paucity of literature on this topic, as well as the breadth of relevant subjects encompassed, prevented us from obtaining the quantity and quality of data required to perform a systematic review or meta-analysis. In particular, the lack of standardized methods to evaluate and mitigate bias precludes any definitive conclusions regarding their suitability in clinical ML applications. However, the scoping review provides a methodological framework for critical evaluation of a previously uncharacterized area of research and draws attention to the lack of standardization regarding racial bias mitigation in clinical ML development. We emphasize the need for further work to build on this important aspect of the medical ML literature.

Conclusion

Algorithmic fairness in clinical ML is a primary concern in its ethical adoption. As medical ML applications continue to approach widespread adoption across a multitude of clinical settings, potential racial biases in ML models must be proactively evaluated and mitigated in order to prevent patient harm and propagation of inequities in health care. The adoption of algorithmic fairness principles in medicine remains nascent, and further research is needed to standardize best practices for fairness evaluation and bias mitigation. We recommend that researchers and journal editors emphasize standardized reporting and data availability in ML studies to improve transparency and facilitate future research. Continued interrogation of biases in clinical ML models is needed to ensure fairness and maximize the benefits of ML in medicine.

Authors' Contributions

No part of this work has been previously published.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary data file containing bias definitions, search strategy, and a table with study data set characteristics.
[PDF File (Adobe PDF File), 215 KB - medinform_v10i5e36388_app1.pdf]

References

1. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
2. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020 Jan 01;577(7788):89-94. [doi: 10.1038/s41586-019-1799-6]
3. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform 2019 Aug 16;7(3):e10010 [FREE Full text] [doi: 10.2196/10010] [Medline: 31420959]
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
5. 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]
6. Obermeyer Z, Emanuel EJ. Predicting the future: big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
7. Domingo J, Galal G, Huang J, Soni P, Mukhin V, Altman C, et al. Preventing delayed and missed care by applying artificial intelligence to trigger radiology imaging follow-up. NEJM Catalyst 2022 Mar 16;3(4). [doi: 10.1056/cat.21.0469]
8. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 2018 Dec 12;18(Suppl 4):122 [FREE Full text] [doi: 10.1186/s12911-018-0677-8] [Medline: 30537977]
9. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019 Jan;25(1):30-36 [FREE Full text] [doi: 10.1038/s41591-018-0307-0] [Medline: 30617336]
10. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digital Health 2020 Dec;2(12):e677-e680. [doi: 10.1016/s2589-7500(20)30200-4]
11. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195 [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
12. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019 Mar 12;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
13. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021 Jul;54(6):1-35. [doi: 10.1145/3457607]
14. Bailey ZD, Feldman JM, Bassett MT. How structural racism works: racist policies as a root cause of U.S. racial health inequities. N Engl J Med 2021 Feb 25;384(8):768-773. [doi: 10.1056/nejmms2025396]
15. Rodriguez JA, Clark CR, Bates DW. Digital health equity as a necessity in the 21st Century Cures Act era. JAMA 2020 Jun 16;323(23):2381-2382. [doi: 10.1001/jama.2020.7858] [Medline: 32463421]
16. Unger JM, Vaidya R, Albain KS, LeBlanc M, Minasian LM, Gotay CC, et al. Sex differences in risk of severe adverse events in patients receiving immunotherapy, targeted therapy, or chemotherapy in cancer clinical trials. JCO 2022 May 01;40(13):1474-1486. [doi: 10.1200/jco.21.02377]
17. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight: reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020 Aug 27;383(9):874-882 [FREE Full text] [doi: 10.1056/NEJMms2004740] [Medline: 32853499]
18. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med 2020 Jul 30;3(1):99 [FREE Full text] [doi: 10.1038/s41746-020-0304-9] [Medline: 32821854]
19. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018 Dec 04;169(12):866. [doi: 10.7326/m18-1990]
20. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021 Jul 20;4(1):123-144. [doi: 10.1146/annurev-biodatasci-092820-114757] [Medline: 34396058]

21. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Front Artif Intell 2020;3:561802 [FREE Full text] [doi: 10.3389/frai.2020.561802] [Medline: 33981989]
22. Wick M, Panda S, Tristan J. Unlocking fairness: a trade-off revisited. 2019 Presented at: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems; December 2019; Vancouver, BC, Canada URL: https://jtristan.github.io/papers/neurips19.pdf
23. Verma S, Rubin J. Fairness definitions explained. 2018 Presented at: Proceedings of the International Workshop on Software Fairness; 2018; Gothenburg, Sweden. [doi: 10.1145/3194770.3194776]
24. Friedler S, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton E, Roth D. A comparative study of fairness-enhancing interventions in machine learning. 2019 Presented at: Proceedings of the Conference on Fairness, Accountability, Transparency; 2019; Atlanta, GA. [doi: 10.1145/3287560.3287589]
25. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. 2017 Presented at: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017); January 2017; Berkeley, CA. [doi: 10.4230/LIPIcs.ITCS.2017.43]
26. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. Lancet Digital Health 2020 May;2(5):e221-e223. [doi: 10.1016/s2589-7500(20)30065-0]
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
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 Sep 04;169(7):467. [doi: 10.7326/M18-0850]
29. Abubakar A, Ugail H, Bukar AM. Assessment of human skin burns: a deep transfer learning approach. J Med Biol Eng 2020 Apr 24;40(3):321-333. [doi: 10.1007/s40846-020-00520-z]
30. Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR Public Health Surveill 2020 Oct 22;6(4):e22400 [FREE Full text] [doi: 10.2196/22400] [Medline: 33090117]
31. Briggs E, Hollmén J. Mitigating discrimination in clinical machine learning decision support using algorithmic processing techniques. In: Appice A, Tsoumakas G, Manolopoulos Y, Matwin S, editors. International Conference on Discovery Science. Vol 12323. Cham: Springer International; 2020:19-33.
32. Burlina P, Joshi N, Paul W, Pacheco KD, Bressler NM. Addressing artificial intelligence bias in retinal diagnostics. Transl Vis Sci Technol 2021 Feb 05;10(2):13 [FREE Full text] [doi: 10.1167/tvst.10.2.13] [Medline: 34003898]
33. Chen I, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics 2019 Feb 01;21(2):E167-E179 [FREE Full text] [doi: 10.1001/amajethics.2019.167] [Medline: 30794127]
34. Gianattasio K, Ciarleglio A, Power M. Development of algorithmic dementia ascertainment for racial/ethnic disparities research in the US Health and Retirement Study. Epidemiology (Cambridge, Mass) 2020;31(1):126-133. [doi: 10.1097/ede.0000000000001101]
35. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. Circ Arrhythm Electrophysiol 2020 Mar;13(3):e007988 [FREE Full text] [doi: 10.1161/CIRCEP.119.007988] [Medline: 32064914]
36. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
37. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA Netw Open 2021 Apr 01;4(4):e213909 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.3909] [Medline: 33856478]
38. Seyyed-Kalantari L, Liu G, McDermott M, Chen I, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pacific Symp Biocomput 2021;26:232-243. [doi: 10.1142/9789811232701_0022]
39. Thompson H, Sharma B, Bhalla S, Boley R, McCluskey C, Dligach D, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. J Am Med Inform Assoc 2021 Oct 12;28(11):2393-2403 [FREE Full text] [doi: 10.1093/jamia/ocab148] [Medline: 34383925]
40. Wissel BD, Greiner HM, Glauser TA, Mangano FT, Santel D, Pestian JP, et al. Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. Epilepsia 2019 Sep 23;60(9):e93-e98 [FREE Full text] [doi: 10.1111/epi.16320] [Medline: 31441044]
41. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. 2016 Presented at: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems; December 2016; Barcelona, Spain.
42. Kamishima T, Akaho S, Asoh H, Sakuma J. Fairness-aware classifier with prejudice remover regularizer. 2012 Presented at: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; September 2012; Berlin, Heidelberg. [doi: 10.1007/978-3-642-33486-3_3]

43. Liu L, Dean S, Rolf E, Simchowitz M, Hardt M. Delayed impact of fair machine learning. 2019 Presented at: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; August 10-16, 2019; Macao. [doi: 10.24963/ijcai.2019/862]
44. Rodolfa K, Salomon E, Haynes L, Mendieta I, Larson J, Ghani R. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. 2020 Presented at: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020; Barcelona, Spain. [doi: 10.1145/3351095.3372863]
45. Kusner MJ, Loftus JR. The long road to fairer algorithms. Nature 2020 Feb 04;578(7793):34-36. [doi: 10.1038/d41586-020-00274-3] [Medline: 32020122]
46. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51. [doi: 10.7326/m18-1376]
47. Alvi M, Zisserman A, Nellåker C. Turning a blind eye: explicit removal of biases and variation from deep neural network embeddings. 2018 Presented at: 15th European Conference on Computer Vision; September 2018; Munich, Germany. [doi: 10.1007/978-3-030-11009-3_34]
48. Vig J, Gehrmann S, Belinkov Y. Investigating gender bias in language models using causal mediation analysis. 2020 Presented at: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems; December 2020; Virtual Conference URL: https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf
49. Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. PLoS One 2020 Dec 17;15(12):e0240376 [FREE Full text] [doi: 10.1371/journal.pone.0240376] [Medline: 33332380]
50. Bolukbasi T, Chang K, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. 2016 Presented at: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems; December 2016; Barcelona, Spain URL: https://papers.nips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf
51. Huang Y, Li W, Macheret F, Gabriel R, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020 Apr 01;27(4):621-633 [FREE Full text] [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
52. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun 2021 Jul 20;12(1):4423 [FREE Full text] [doi: 10.1038/s41467-021-24698-1] [Medline: 34285218]
53. Bellamy RKE, Dey K, Hind M, Hoffman S, Houde S, Kannan K, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019 Jul 1;63(4/5):4:1-4:15. [doi: 10.1147/jrd.2019.2942287]
54. Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viegas F, Wilson J. The what-if tool: interactive probing of machine learning models. IEEE Trans Visual Comput Graphics 2020;26(1):56-65. [doi: 10.1109/tvcg.2019.2934619]
55. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019 May 13;1(5):206-215. [doi: 10.1038/s42256-019-0048-x]
56. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learning Syst 2021 Nov;32(11):4793-4813. [doi: 10.1109/tnnls.2020.3027314]
57. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med 2021 Mar 24;13(586):eabb1655. [doi: 10.1126/scitranslmed.abb1655] [Medline: 33762434]
58. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021 Mar 15;3(3):199-217. [doi: 10.1038/s42256-021-00307-0]
59. Huang J, Shlobin N, DeCuypere M, Lam S. Deep learning for outcome prediction in neurosurgery: a systematic review of design, reporting, and reproducibility. Neurosurgery 2022;90(1):16-38. [doi: 10.1227/neu.0000000000001736]
60. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 2020 Jul 28;10(1):12598 [FREE Full text] [doi: 10.1038/s41598-020-69250-1] [Medline: 32724046]
61. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inform 2021 Apr 28;28(1):e100289 [FREE Full text] [doi: 10.1136/bmjhci-2020-100289] [Medline: 33910923]
62. Bozkurt S, Cahan E, Seneviratne M, Sun R, Lossio-Ventura JA, Ioannidis JPA, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. J Am Med Inform Assoc 2020 Dec 09;27(12):1878-1884 [FREE Full text] [doi: 10.1093/jamia/ocaa164] [Medline: 32935131]
63. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-a scoping reviewA scoping review. J Am Acad Dermatol 2021 Jul 10. [doi: 10.1016/j.jaad.2021.06.884] [Medline: 34252465]

64. Hernandez-Boussard T, Bozkurt S, Ioannidis J, Shah N. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc 2020 Dec 09;27(12):2011-2015 [FREE Full text] [doi: 10.1093/jamia/ocaa088] [Medline: 32594179]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
FNR: false-negative rate
FPR: false-positive rate
LASSO: least absolute shrinkage and selection operator
ML: machine learning
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses
TPR: true-positive rate

Edited by C Lovis; submitted 12.01.22; peer-reviewed by H Turbe; comments to author 13.02.22; revised version received 17.02.22; accepted 27.03.22; published 31.05.22.

Please cite as:
Huang J, Galal G, Etemadi M, Vaidyanathan M
Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review
JMIR Med Inform 2022;10(5):e36388
URL: https://medinform.jmir.org/2022/5/e36388
doi:10.2196/36388
PMID:35639450

Original Paper

Web-Based Software Tools for Systematic Literature Review in Medicine: Systematic Search and Feature Analysis

Kathryn Cowie[1], BS; Asad Rahmatullah[1], BS; Nicole Hardy[1], MSc; Karl Holub[1], BS; Kevin Kallmes[1], MA, JD

Nested Knowledge, Saint Paul, MN, United States

Corresponding Author:
Kevin Kallmes, MA, JD
Nested Knowledge
1430 Avon St. N.
Saint Paul, MN, 55117
United States
Phone: 1 5072717051
Email: kevinkallmes@supedit.com

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2022/11/e43520

Abstract

Background: Systematic reviews (SRs) are central to evaluating therapies but have high costs in terms of both time and money. Many software tools exist to assist with SRs, but most tools do not support the full process, and transparency and replicability of SR depend on performing and presenting evidence according to established best practices.

Objective: This study aims to provide a basis for comparing and selecting between web-based software tools that support SR, by conducting a feature-by-feature comparison of SR tools.

Methods: We searched for SR tools by reviewing any such tool listed in the SR Toolbox, previous reviews of SR tools, and qualitative Google searching. We included all SR tools that were currently functional and required no coding, and excluded reference managers, desktop applications, and statistical software. The list of features to assess was populated by combining all features assessed in 4 previous reviews of SR tools; we also added 5 features (manual addition, screening automation, dual extraction, living review, and public outputs) that were independently noted as best practices or enhancements of transparency and replicability. Then, 2 reviewers assigned binary *present* or *absent* assessments to all SR tools with respect to all features, and a third reviewer adjudicated all disagreements.

Results: Of the 53 SR tools found, 55% (29/53) were excluded, leaving 45% (24/53) for assessment. In total, 30 features were assessed across 6 classes, and the interobserver agreement was 86.46%. Giotto Compliance (27/30, 90%), DistillerSR (26/30, 87%), and Nested Knowledge (26/30, 87%) support the most features, followed by EPPI-Reviewer Web (25/30, 83%), LitStream (23/30, 77%), JBI SUMARI (21/30, 70%), and SRDB.PRO (VTS Software) (21/30, 70%). Fewer than half of all the features assessed are supported by 7 tools: RobotAnalyst (National Centre for Text Mining), SRDR (Agency for Healthcare Research and Quality), SyRF (Systematic Review Facility), Data Abstraction Assistant (Center for Evidence Synthesis in Health), SR Accelerator (Institute for Evidence-Based Healthcare), RobotReviewer (RobotReviewer), and COVID-NMA (COVID-NMA). Notably, of the 24 tools, only 10 (42%) support direct search, only 7 (29%) offer dual extraction, and only 13 (54%) offer living/updatable reviews.

Conclusions: DistillerSR, Nested Knowledge, and EPPI-Reviewer Web each offer a high density of SR-focused web-based tools. By transparent comparison and discussion regarding SR tool functionality, the medical community can both choose among existing software offerings and note the areas of growth needed, most notably in the support of living reviews.

(JMIR Med Inform 2022;10(5):e33219) doi:10.2196/33219

KEYWORDS

software tools; feature analysis; systematic reviews

Introduction

Systematic Review Costs and Gaps

According to the Centre for Evidence-Based Medicine, systematic reviews (SRs) of high-quality primary studies represent the highest level of evidence for evaluating therapeutic performance [1]. However, although vital to evidence-based medical practice, SRs are time-intensive, taking an average of 67.3 weeks to complete [2] and costing leading research institutions over US $141,000 in labor per published review [3]. Owing to the high costs in researcher time and complexity, up-to-date reviews cover only 10% to 17% of primary evidence in a representative analysis of the lung cancer literature [4]. Although many qualitative and noncomprehensive publications provide some level of summative evidence, SRs—defined as reviews of "evidence on a clearly formulated question that use systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyze data from the studies that are included" [5]—are distinguished by both their structured approach to finding, filtering, and extracting from underlying articles and the resulting comprehensiveness in answering a concrete medical question.

Software Tools for Systematic Review

Software tools that assist with central SR activities—retrieval (searching or importing records), appraisal (screening of records), synthesis (content extraction from underlying studies), and documentation/output (presentation of SR outputs)—have shown promise in reducing the amount of effort needed in a given review [6]. Because of the time savings of web-based software tools, institutions and individual researchers engaged in evidence synthesis may benefit from using these tools in the review process [7].

Existing Studies of Software Tools

However, choosing among the existing software tools presents a further challenge to researchers; in the SR Toolbox [8], there are >240 tools indexed, of which 224 support health care reviews. Vitally, few of these tools can be used for each of the steps of SR, so comparing the features available through each tool can assist researchers in selecting an SR tool to use. This selection can be informed by feature analysis; for example, a previously published feature analysis compared 15 SR tools [9] across 21 subfeatures of interest and found that DistillerSR (Evidence Partners), EPPI-Reviewer (EPPI-Centre), SWIFT-Active Screener (Sciome), and Covidence (Cochrane) support the greatest number of features as of 2019. Harrison et al [10], Marshall et al [11], and Kohl et al [12] have completed similar analyses, but each feature assessment selected a different set of features and used different qualitative feature assessment methods, and none covered all SR tools currently available.

The SR tool landscape continues to evolve; as existing tools are updated, new software is made available to researchers, and new feature classes are developed. For instance, despite the growth of calls for living SRs, that is, reviews where the outputs are updated as new primary evidence becomes available, no feature analysis has yet covered this novel capability. Furthermore, the leading feature analyses [9-12] have focused on the screening phase of review, meaning that no comparison of data extraction capabilities has yet been published.

Feature Analysis of Systematic Review Tools

The authors, who are also the developers of the Nested Knowledge platform for SR and meta-analysis (Nested Knowledge, Inc) [13], have noted the lack of SR feature comparison among new tools and across all feature classes (retrieval, appraisal, synthesis, documentation/output, administration of reviews, and access/support features). To provide an updated feature analysis comparing SR software tools, we performed a feature analysis covering the full life cycle of SR across software tools.

Methods

Search Strategy

We searched the SR tools for assessment in 3 ways: first, we identified any SR tool that was published in existing reviews of SR tools (Table S1 in Multimedia Appendix 1). Second, we reviewed SR Toolbox [8], a repository of indexed software tools that support the SR process. Third, we performed a Google search for *Systematic review software* and identified any software tool that was among the first 5 pages of results. Furthermore, for any library resource pages that were among the search results, we included any SR tools mentioned by the library resource page that met our inclusion criteria. The search was completed between June and August 2021. Four additional tools, namely SRDR+ (Agency for Healthcare Research and Quality), Systematic Review Assistant-Deduplication Module (Institute for Evidence-Based Healthcare), Giotto Compliance, and Robotsearch (Robotsearch), were assessed in December 2021 following reviewer feedback.

Selection of Software Tools

The inclusion and exclusion criteria were determined by 3 authors (KK, KH, and KC). Among our search results, we queued up all software tools that had descriptions meeting our inclusion criteria for full examination of the software in a second round of review. We included any that were functioning web-based tools that require no coding by the user to install or operate, so long as they were used to support the SR process and can be used to review clinical or preclinical literature. The *no coding* requirement was established because the target audience of this review is medical researchers who are selecting a review software to use; thus, we aim to review only tools that this broad audience is likely to be able to adopt. We also excluded desktop applications, statistical packages, and tools built for reviewing software engineering and social sciences literature, as well as reference managers, to avoid unfairly casting these tools as incomplete review tools (as they would each score quite low in features that are not related to reference management). All software tools were screened by one reviewer (KC), and inclusion decisions were reviewed by a second (KK).

Selection of Features of Interest

We built on the previous comparisons of SR tools published by Van der Mierden et al [9], Harrison et al [10], Marshall et al [11], and Kohl et al [12], which assign features a level of

importance and evaluate each feature in reference screening tools. As the studies by Van der Mierden et al [9] and Harrison et al [10] focus on reference screening, we supplemented the features with features identified in related reviews of SR tools (Table S1 in Multimedia Appendix 1). From a study by Kohl et al [12], we added database search, risk of bias assessment (critical appraisal), and data visualization. From Marshall et al [11], we added report writing.

We added 4 more features based on their importance to software-based SR: manual addition of records, automated full-text retrieval, dual extraction of studies, risk of bias (critical appraisal), living SR, and public outputs. Each addition represents either a best practice in SR [14] or a key feature for the accuracy, replicability, and transparency of SR. Thus, in total, we assessed the presence or absence of 30 features across 6 categories: retrieval, appraisal, synthesis, documentation/output, administration/project management, and access/support.

We adopted each feature unless it was outside of the SR process, it was required for inclusion in the present review, it duplicated another feature, it was not a discrete step for comparison, it was not necessary for English language reviews, it was not necessary for a web-based software, or it related to reference management (as we excluded reference managers from the present review). Table 1 shows all features not assessed, with rationale.

Table 1. Features from systematic reviews not assessed in this review, with rationale.

Features not assessed	Rationale
Functional	Part of our inclusion criteria
Reference allocation	Reference management excluded from this review
Randomizing order of references	Not part of systematic review process
Non-Latin character support	Review focused on English language systematic review software
Straightforward system requirements	Part of our inclusion criteria
Installation guide	Not necessary for web-based software
No coding	Part of our inclusion criteria
Mobile- or tablet-responsive interface	Not necessary for web-based software
Other stages	Not a discrete or comparable step
Multiple projects	Not part of the systematic review process
Work allocation	Duplicated with “distinct user roles”
Export of decisions	Duplicated with export
User setup	Duplicated with “distinct user roles”
Filter references	Duplicated with screening records
Search references	Duplicated with “database search”
Insecure website	Information not available to reviewers
Security	Information not available to reviewers
Setting up review	Not a discrete or comparable step
Automated analysis	Not a discrete or comparable step
Text analysis	Not part of the systematic review process
Report validation	Not part of the systematic review process
Document management	Reference management excluded from this review
Bibliography	Reference management excluded from this review

Feature Assessment

To minimize bias concerning the subjective assessment of the necessity or desirability of features or of the relative performance of features, we used a binary assessment where each SR tool was scored 0 if a given feature was not present or 1 if a feature was present. Tools were assessed between June and August 2021. We assessed 30 features, divided into 6 feature classes. Of the 30 features, 77% (23/30) were identified in existing literature, and 23% (7/30) were added by the authors (Table 2).

Table 2. The criteria for each selected feature, as well as the rationale.

Classification and variable name and coding		Feature from	Rationale (if added by authors)
Retrieval			
Database search	1—literature search through API[a] Integration with a database; 0—no method for retrieving studies directly from a database	Kohl et al [12], Marshall et al [11]	—[b]
Reference importing	1—import of references as RIS[c] files or other file types; 0—references have to be entered manually	Harrison et al [10], Van der Mierden et al [9]	—
Manual addition	1—add a reference by entering study metadata; 0—no method for adding individual references and gray literature	Added by the authors	Ability to add expert additions is called for by the PRISMA[d] 2020 guidelines and checklist [14]
Attaching full-text PDFs	1—ability to import or upload full-text PDFs associated with each study under review; 0—no method for importing full-text PDFs in the screening process	Harrison et al [10], Van der Mierden et al [9]	—
Automated full-text retrieval	1—ability to fetch some or all full texts via API or other nonmanual method; 0—full texts must be uploaded manually, or full-text upload not supported	Added by the authors	Full texts are required for content extraction, and manual upload represents a major time investment by the user
Appraisal			
Title/abstract screening	1—inclusion and exclusion by title and abstract only; 0—no system for inclusion and exclusion of references by title and abstract	Harrison et al [10], Van der Mierden et al [9]	—
Full-text screening	1—a distinct full-text screening phase; 0—there is no full-text screening phase	Harrison et al [10], Van der Mierden et al [9]	—
Dual screening and adjudication	1—choice for single or double screening and a method for resolving conflicts; 0—no ability to configure screening mode or no ability to resolve conflicts	Harrison et al [10], Van der Mierden et al [9]	—
Keyword highlighting	1—abstract keywords are highlighted. Keywords can be user or AI[e]-determined; 0—No keyword highlighting is possible	Harrison et al [10], Van der Mierden et al [9]	—
Machine learning/automation (screening)	1—has a form of machine learning or automation of the screening process; 0—does not support any form of machine learning or automation of the screening process	Added by the authors	Automated screening has been called for by the scientific community [15]
Deduplication of references	1—automatically identifies duplicate references or marks potential duplicates for manual review; 0—has no mechanism for deduplication	Harrison et al [10], Kohl et al [12]	—
Extraction			
Tagging references	1—ability to attach tags that reflect the content of underlying studies to specific references; 0—no means for attaching content-related tags to references	Van der Mierden et al [9], Kohl et al [12]	—
Data extraction	1—facilitates extraction and storage of quantitative data into a form or template; 0—does not permit extraction and storage or quantitative data	Harrison et al [10], Kohl et al [12], Marshall et al [11]	—
Dual extraction	1—ability for 2 independent reviewers to collect on each study and for a third person to adjudicate differences; 0—no ability to have independent extraction and adjudication	Added by the authors	Dual extraction improves the accuracy of data gathering [16]
Risk of bias	1—supports critical appraisal of studies through risk of bias assessments; 0—no built-in features or templates to assess risk of bias	Kohl et al [12]	—
Documentation/output			
Flow diagram creation	1—automated or semiautomated creation of PRISMA flow diagrams; 0—the tool cannot automatically provide a flow diagram meeting the PRISMA criteria	Van der Mierden et al [9]	—

Classification and variable name and coding		Feature from	Rationale (if added by authors)
Manuscript writing	1—ability to write or edit a report or manuscript; 0—no ability to write or edit a report or manuscript	Marshall et al [11]	—
Citation management	1—ability to insert citations based on stored study metadata into a text editor; 0—no ability to insert citations into a document	Added by the authors	The ability to add and manage citations is necessary to document the source of review data
Data visualizations	1—generation of figures or tables to assist with data presentation; 0—no built-in way to generate figures or tables	Kohl et al [12]	—
Export	1—supports export of references, study metadata, or collected data; 0—has no export feature	Harrison et al [10], Van der Mierden et al [9]	—
Admin			
Protocol	1—supports protocol development or filling in a research question template; 0—no protocol development or templates	Kohl et al [12], Marshall et al [11]	—
Distinct user roles	1—distinct user roles and permissions; 0—no distinct roles; everybody has the same role and rights in the project	Harrison et al [10], Van der Mierden et al [9], Marshall et al [11]	—
Activity monitoring	1—software monitors and displays progress through the project; 0—there is no way to determine overall progress of the project (eg, % completed)	Harrison et al [10], Van der Mierden et al [9]	—
Comments or chat	1—ability to leave comments or notes on studies; 0—it is not possible to attach comments to references	Van der Mierden et al [9]	—
Training	1—there are publicly available web-based tutorials, help pages, training videos, or forums maintained by the software provider; 0—there are no accessible tutorials or training materials maintained by the software provider	Harrison et al [10], Marshall et al {11]	—
Customer support	1—customer support, such as support contact information, is provided on request; 0—customer support is not clearly available	Van der Mierden et al [9]	—
Access and support			
Pricing (free to use)	1—a free version is available for users; 0—the tool must be purchased, or free or trial accounts have severe limitations that can compromise the systematic review	Harrison et al [10], Van der Mierden et al [9], Marshall et al [11]	—
Living/updatable	1—new records can be added after a project has been completed; 0—new records cannot be added after a project has been completed	Added by the authors	Living systematic review has been called for as a novel paradigm solving the main limitation of systematic review [17]
Public outputs	1—web-based visualizations or writing can be made publicly visible; 0—review data and outputs cannot be made publicly visible	Added by the authors	Web-based availability of systematic review outputs is important for transparency and replicability of research [18]
User collaboration	1—multiple users can work simultaneously on 1 review; 0—it is not possible for multiple users to work at the same time on the same project, independently	Harrison et al [10], Van der Mierden et al [9], Marshall et al [11]	—

[a]API: application programming interface.

[b]Rationale only provided for features added in this review; all other features were drawn from existing feature analyses of Systematic Review Software Tools.

[c]RIS: Research Information System.

[d]PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

[e]AI: artificial intelligence.

Evaluation of Tools

For tools with free versions available, each of the researchers created an account and tested the program to determine feature presence. We also referred to user guides, publications, and training tutorials. For proprietary software, we gathered information on feature offerings from marketing webpages, training materials, and video tutorials. We also contacted all proprietary software providers to give them the opportunity to comment on feature offerings that may have been left out of

those materials. Of the 8 proprietary software providers contacted, 38% (3/8) did not respond, 50% (4/8) provided feedback on feature offerings, and 13% (1/8) declined to comment. When providers provided feedback, we re-reviewed the features in question and altered the assessment as appropriate. One provider gave feedback after initial puplication, prompting issuance of a correction.

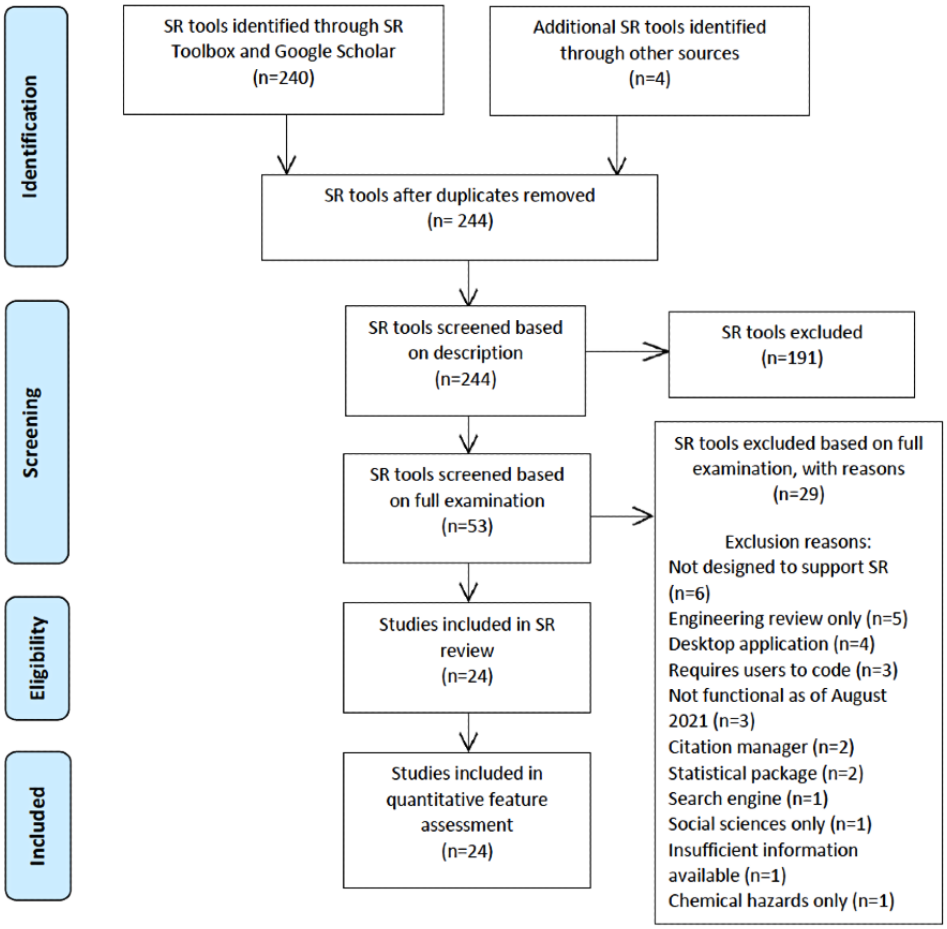
Feature assessment was completed independently by 2 reviewers (KC and AR), and all disagreements were adjudicated by a third (KK). Interobserver agreement was calculated using standard methods [19] as applied to binary assessments. First, the 2 independent assessments were compared, and the number of disagreements was counted per feature, per software. For each feature, the total number of disagreements was counted and divided by the number of software tools assessed. This provided a per-feature variability percentage; these percentages were averaged across all features to provide a cumulative interobserver agreement percentage.

Results

Identification of SR Tools

We reviewed all 240 software tools offered on SR Toolbox and sent forward all studies that, based on the software descriptions, could meet our inclusion criteria; we then added in all software tools found on Google Scholar. This strategy yielded 53 software tools that were reviewed in full (Figure 1 shows the PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses]-based chart). Of these 53 software tools, 55% (29/53) were excluded. Of the 29 excluded tools, 17% (5/29) were built to review software engineering literature, 10% (3/29) were not functional as of August 2021, 7% (2/29) were citation managers, and 7% (2/29) were statistical packages. Other excluded tools included tools not designed for SRs (6/29, 21%), desktop applications (4/29, 14%), tools requiring users to code (3/29, 10%), a search engine (1/29, 3%), and a social science literature review tool (1/29, 3%). One tool, Research Screener [20], was excluded owing to insufficient information available on supported features. Another tool, the Health Assessment Workspace Collaborative, was excluded because it is designed to assess chemical hazards.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)-based chart showing the sources of all tools considered for inclusion, including 2-phase screening and reasons for all exclusions made at the full software review stage. SR: systematic review.



Overview of SR Tools

We assessed the presence of features in 24 software tools, of which 71% (17/24) are designed for health care or biomedical sciences. In addition, 63% (15/24) of the analyzed tools support the full SR process, meaning they enable search, screening, extraction, and export, as these are the basic capabilities necessary to complete a review in a single software tool. Furthermore, 21% (5/34) of the tools support the screening stage (Table 3).

Table 3. Breakdown of software tools for systematic review by process type (full process, screening, extraction, or visualization; n=24).

Type	Tools, n (%)	Software tools
Full process	15 (63)	Cadima, Covidence, Colandr, DistillerSR, EPPI-Reviewer Web, Giotto Compliance, JBI SUMARI, LitStream, Nested Knowledge, PICOPortal, Revman Web, SRDB.PRO, SRDR+, SyRF, SysRev
Screening	5 (21)	Abstrackr, Rayyan, RobotAnalyst, SWIFT-Active Screener, SR Accelerator
Extraction	3 (13)	Data Abstraction Assistant, RobotReviewer, SRDR
Visualization	1 (4)	COVID-NMA

Data Gathering

Interobserver agreement between the 2 reviewers gathering data features was 86.46%, meaning that across all feature assessments, the 2 reviewers disagreed on <15% of the applications. Final assessments are summarized in Table 4, and Table S2 in Multimedia Appendix 2 shows the interobserver agreement on a per–SR tool and per-feature basis. Interobserver agreement was ≥70% for every feature assessed and for all SR tools except 3: LitStream (ICF; 53.3%), RevMan Web (Cochrane; 50%), and SR Accelerator (Institute for Evidence-Based Healthcare; 53.3%); on investigation, these low rates of agreement were found to be due to name changes and versioning (LitStream and RevMan Web) and due to the modular nature of the subsidiary offerings (SR Accelerator). An interactive, updatable visualization of the features offered by each tool is available in the Systematic Review Methodologies Qualitative Synthesis.

Table 4. Feature assessment scores by feature class for each systematic review tool analyzed. The total number of features across all feature classes is presented in descending order.

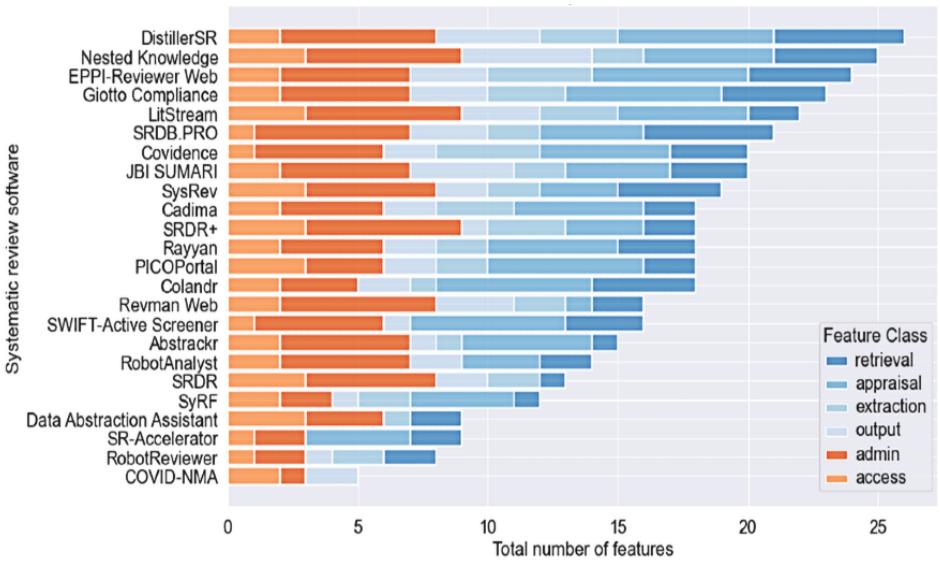
Systematic review tool	Retrieval (n=5), n (%)	Appraisal (n=6), n (%)	Extraction (n=4), n (%)	Output (n=5), n (%)	Admin (n=6), n (%)	Access (n=4), n (%)	Total (n=30), n (%)
Giotto Compliance	5 (100)	6 (100)	4 (100)	3 (60)	6 (100)	3 (75)	27 (90)
DistillerSR	5 (100)	6 (100)	3 (75)	4 (80)	6 (100)	2 (50)	26 (87)
Nested Knowledge	4 (80)	5 (83)	2 (50)	5 (100)	6 (100)	4 (100)	26 (87)
EPPI-Reviewer Web	4 (80)	6 (100)	4 (100)	3 (60)	5 (83)	3 (75)	25 (83)
LitStream	2 (40)	5 (83)	3 (75)	3 (60)	6 (100)	4 (100)	23 (77)
JBI SUMARI	3 (60)	4 (67)	2 (50)	4 (80)	5 (83)	3 (75)	21 (70)
SRDB.PRO	5 (100)	4 (67)	2 (50)	3 (60)	6 (100)	1 (25)	21 (70)
Covidence	3 (60)	5 (83)	4 (100)	2 (40)	5 (83)	1 (25)	20 (67)
SysRev	4 (80)	3 (50)	2 (50)	2 (40)	5 (83)	4 (100)	20 (67)
Cadima	2 (40)	5 (83)	3 (75)	2 (40)	4 (67)	3 (75)	19 (63)
SRDR+	2 (40)	3 (50)	3 (75)	1 (20)	6 (100)	4 (100)	19 (63)
Colandr	4 (80)	6 (100)	1 (25)	2 (40)	3 (50)	2 (50)	18 (60)
PICOPortal	2 (40)	6 (100)	2 (50)	2 (40)	3 (50)	3 (75)	18 (60)
Rayyan	3 (60)	5 (83)	2 (50)	2 (40)	4 (50)	2 (50)	18 (60)
Revman Web	2 (40)	1 (17)	2 (50)	3 (60)	6 (100)	3 (75)	17 (57)
SWIFT-Active Screener	3 (60)	6 (100)	0 (0)	1 (20)	5 (83)	1 (25)	16 (53)
Abstrackr	1 (20)	5 (83)	1 (25)	1 (20)	5 (83)	2 (50)	15 (50)
RobotAnalyst	2 (40)	3 (50)	0 (0)	2 (40)	5 (83)	2 (50)	14 (47)
SRDR	1 (20)	0 (0)	2 (50)	2 (40)	5 (83)	4 (100)	14 (47)
SyRF	1 (20)	4 (67)	2 (50)	1 (20)	2 (33)	2 (50)	12 (40)
Data Abstraction Assistant	2 (40)	0 (0)	1 (25)	0 (0)	3 (50)	4 (100)	10 (33)
SR-Accelerator	2 (40)	4 (67)	0 (0)	0 (0)	2 (33)	1 (25)	9 (30)
RobotReviewer	2 (40)	0 (0)	2 (50)	1 (20)	2 (33)	1 (25)	8 (27)
COVID-NMA	0 (0)	0 (0)	0 (0)	2 (40)	1 (17)	3 (75)	6 (20)

Feature Assessment

Giotto Compliance (27/30, 90%), DistillerSR (26/30, 87%), and Nested Knowledge (26/30, 87%) support the most features, followed by EPPI-Reviewer Web (25/30, 83%), LitStream (23/30, 77%), JBI SUMARI (21/30, 70%), and SRDB.PRO (VTS Software) (21/30, 70%).

The top 16 software tools are ranked by percent of features from highest to lowest in Figure 2. Fewer than half of all features are supported by 7 tools: RobotAnalyst (National Centre for Text Mining), SRDR (Agency for Healthcare Research and Quality), SyRF (Systematic Review Facility), Data Abstraction Assistant (Center for Evidence Synthesis in Health, Institute for Evidence-Based Healthcare), SR-Accelerator, RobotReviewer (RobotReviewer), and COVID-NMA (COVID-NMA; Table 3).

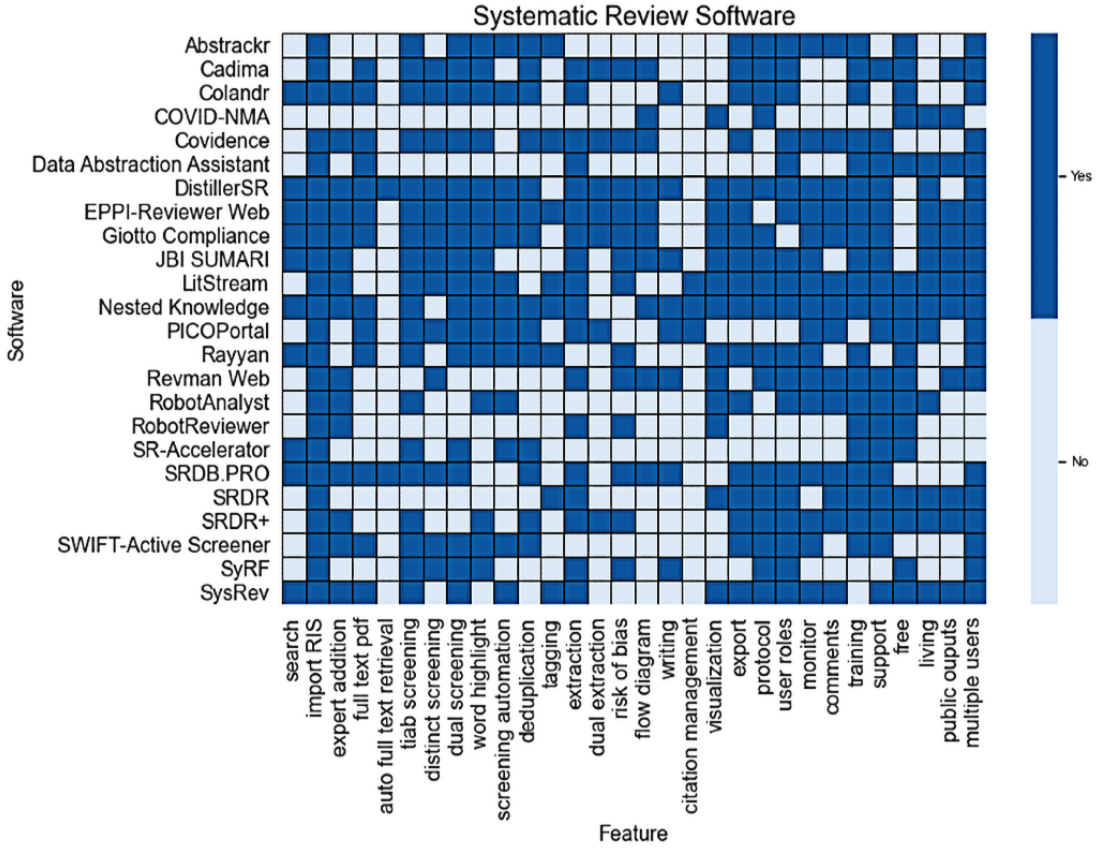
Figure 2. Stacked bar chart comparing the percentage of supported features, broken down by their feature class (retrieval, appraisal, extraction, output, admin, and access), among all analyzed software tools.



Feature Assessment: Breakout by Feature Class

Of all 6 feature classes, administrative features are the most supported, and output and extraction features are the least supported (Figure 3). Only 3 tools, Covidence (Cochrane), EPPI-Reviewer, and Giotto Compliance, offer all 4 extraction features (Table 4). DistillerSR and Giotto support all 5 retrieval features, while Nested Knowledge supports all 5 documentation/output features. Colandr, DistillerSR, EPPI-Reviewer, Giotto Compliance, and PICOPortal support all 6 appraisal features.

Figure 3. Heat map of features observed in 24 analyzed software tools. Dark blue indicates that a feature is present, and light blue indicates that a feature is not present.



Feature Class 1: Retrieval

The ability to search directly within the SR tool was only present for 42% (10/24) of the software tools, meaning that for all other SR tools, the user is required to search externally and import records. The only SR tool that did not enable importing of records was COVID-NMA, which supplies studies directly from the providers of the tool but does not enable the user to do so.

Feature Class 2: Appraisal

Among the 19 tools that have title/abstract screening, all tools except for RobotAnalyst and SRDR+ enable dual screening and adjudication. Reference deduplication is less widespread, with 58% (14/24) of the tools supporting it. A form of machine learning/automation during the screening stage is present in 54% (13/24) of the tools.

Feature Class 3: Extraction

Although 75% (18/24) of the tools offer data extraction, only 29% (7/24) offer dual data extraction (Giotto Compliance, DistillerSR, SRDR+, Cadima [Cadima], Covidence, EPPI-Reviewer, and PICOPortal [PICOPortal]). A total of 54% (13/24) of the tools enable risk of bias assessments.

Feature Class 4: Output

Exporting references or collected data is available in 71% (17/24) of the tools. Of the 24 tools, 54% (13/24) generate figures or tables, 42% (10/24) of tools generate PRISMA flow diagrams, 32% (8/24%) have report writing, and only 13% (3/34) have in-text citations.

Feature Class 5: Admin

Protocols, customer support, and training materials are available in 71% (17/24), 79% (19/24), and 83% (20/24) of the tools, respectively. Of all administrative features, the least well developed are progress/activity monitoring, which is offered 67% (16/24) of the tools, and comments, which are available in 58% (14/24) of the tools.

Feature Class 6: Access

Access features cover both collaboration during the review, cost, and availability of outputs. Of the 24 software tools, 83% (20/24) permit collaboration by allowing multiple users to work on a project. COVID-NMA, RobotAnalyst, RobotReviewer, and SR-Accelerator do not allow multiple users. In addition, of the 24 tools, 71% (17/24) offer a free subscription, whereas 29% (7/24) require paid subscriptions or licenses (Covidence, DistillerSR, EPPI-Reviewer Web, Giotto Compliance, JBI Sumari, SRDB.PRO, and SWIFT-Active Screener). Only 54% (13/24) of the software tools support living, updatable reviews.

Discussion

Principal Findings

Our review found a wide range of options in the SR software space; however, among these tools, many lacked features that are either crucial to the completion of a review or recommended as best practices. Only 63% (15/24) of the SR tools covered the full process from search/import through to extraction and export. Among these 15 tools, only 67% (10/15) had a search functionality directly built in, and only 47% (7/15) offered dual data extraction (which is the gold standard in quality control). Notable strengths across the field include collaborative mechanisms (offered by 20/24, 83% tools) and easy, free access (17/24, 71% of tools are free). Indeed, the top 4 software tools in terms of number of features offered (Giotto Compliance, DistillerSR, Nested Knowledge, and EPPI-Reviewer all offered between 83% and 90% of the features assessed. However, major remaining gaps include a lack of automation of any step other than screening (automated screening offered by 13/24, 54% of tools) and underprovision of living, updatable outputs.

Major Gaps in the Provision of SR Tools

Search

Marshall et al [11] have previously noted that "the user should be able to perform an automated search from within the tool which should identify duplicate papers and handle them accordingly" [11]. Less than a third of tools (7/24, 29%) support search, reference import, and manual reference addition.

Study Selection

Screening of references is the most commonly offered feature and has the strongest offerings across features. All software tools that offer screening also support dual screening (with the exception of RobotAnalyst and SRDR+). This demonstrates adherence to SR best practices during the screening stage.

Automation and Machine Learning

Automation in medical SR screening has been growing. Some form of machine learning or other automation for screening literature is present in over half (13/24, 54%) of all the tools analyzed. Machine learning/screening includes reordering references, topic modeling, and predicting inclusion rates.

Data Extraction

In contrast to screening, extraction is underdeveloped. Although extraction is offered by 75% (18/24) tools, few tools adhere to SR best practices of dual extraction. This is a deep problem in the methods of review, as the error rate for manual extraction without dual extraction is highly variable and has even reached 50% in independent tests [16].

Although single extraction continues to be the only commonly offered method, the scientific community has noted that automating extraction would have value in both time savings and improved accuracy, but the field is as of yet underdeveloped. To quote a recent review on the subject of automated extraction, "[automation] techniques have not been fully utilized to fully or even partially automate the data extraction step of systematic review" [21]. The technologies to automate extraction have not achieved partial extraction at a sufficiently high accuracy level to be adopted; therefore, dual extraction is a pressing software requirement that is unlikely to be surpassed in the near future.

Project Management

Administrative features are well supported by SR software. However, there is a need for improved monitoring of review progress. Project monitoring is offered by 67% (16/24) of the tools, which is among the lowest of all admin features and likely the feature most closely associated with the quality of the

outputs. As collaborative access is common and highly prized, SR software providers should recognize the barriers to collaboration in medical research; lack of mutual awareness, inertia in communication, and time management and capacity constraints are among the leading reasons for failure in interinstitutional research [22]. Project monitoring tools could assist with each of these pain points and improve the transparency and accountability within the research team.

Living Reviews

The scientific community has made consistent demands for SR processes to be rendered updatable, with the goal of improving the quality of evidence available to clinicians, health policymakers, and the medical public [23,24]. Despite these ongoing calls for change, living, updatable reviews are not yet standard in SR software tools. Only 54% (13/24) of the tools support living reviews, largely because living review depends on providing updatability at each step up through to outputs. However, until greater provision of living review tools is achieved, reviews will continue to fall out of date and out of sync with clinical practice [24].

Study Limitations

In our study design, we elected to use a binary assessment, which limited the bias induced by the subjective appeal of any given tool. Therefore, these assessments did not include any comparison of quality or usability among the SR tools. This also meant that we did not use the Desmet [25] method, which ranks features by level of importance. We also excluded certain assessments that may impact user choices such as language translation features or translated training documentation, which is supported by some technologies, including DistillerSR. We completed the review in August 2021 but added several software tools following reviewer feedback; by adding *expert additions* without repeating the entire search strategy, we may have missed SR tools that launched between August and December 2021. Finally, the authors of this study are the designers of one of the leading SR tools, Nested Knowledge, which may have led to tacit bias toward this tool as part of the comparison.

By assessing features offered by web-based SR applications, we have identified gaps in current technologies and areas in need of development. Feature count does not equate to value or usability; it fails to capture benefits of simple platforms, such as ease of use, effective user interface, alignment with established workflows, or relative costs. The authors make no claim about superiority of software based on feature prevalence.

Future Directions

We invite and encourage independent researchers to assess the landscape of SR tools and build on this review. We expect the list of features to be assessed will evolve as research changes. For example, this review did not include features such as the ability to search included studies, reuse of extracted data, and application programming interface calls to read data, which may grow in importance. Furthermore, this review assessed the presence of automation at a high level without evaluating details. A future direction might be characterizing specific types of automation models used in screening, as well as in other stages, for software applications that support SR of biomedical research.

Conclusions

The highest-performing SR tools were DistillerSR, EPPI-Reviewer Web, and Nested Knowledge, each of which offer >80% of features. The most commonly offered and robust feature class was screening, whereas extraction (especially quality-controlled dual extraction) was underprovided. Living reviews, although strongly advocated for in the scientific community, were similarly underprovided by the SR tools reviewed here. This review enables the medical community to complete transparent and comprehensive comparison of SR tools and may also be used to identify gaps in technology for further development by the providers of these or novel SR tools.

Disclaimer

This review of web-based software review software tools represents an attempt to best capture information from software providers' websites, free trials, peer-reviewed publications, training materials, or software tutorials. The review is based primarily on publicly available information and may not accurately reflect feature offerings, as relevant information was not always available or clear to interpret. This evaluation does not represent the views or opinions of any of the software developers or service providers, except those of the authors. The review was completed in August 2021, and readers should refer to the respective software providers' websites to obtain updated information on feature offerings.

Acknowledgments

The authors acknowledge the software development team from Nested Knowledge, Stephen Mead, Jeffrey Johnson, and Darian Lehmann-Plantenberg for their input in designing Nested Knowledge. The authors thank the independent software providers who provided feedback on our feature assessment, which increased the quality and accuracy of the results.

Authors' Contributions

All authors participated in the conception, drafting, and editing of the manuscript.

Conflicts of Interest

KC, NH, and KH work for and hold equity in Nested Knowledge, which provides a software application included in this assessment. AR worked for Nested Knowledge. KL works for and holds equity in Nested Knowledge, Inc, and holds equity in Superior Medical Experts, Inc. KK works for and holds equity in Nested Knowledge, and holds equity in Superior Medical Experts.

Multimedia Appendix 1
Supplementary Table 1: Screening Decisions for SR (systematic review) Tools Reviewed in Full.
[DOCX File , 19 KB - medinform_v10i4e33219_app1.docx]

Multimedia Appendix 2
Supplementary Table 2: Inter-observer Agreement across (1) Systematic Review (SR) Tools and (2) Features Assessed.
[DOCX File , 16 KB - medinform_v10i4e33219_app2.docx]

References

1. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg 2011 Jul;128(1):305-310 [FREE Full text] [doi: 10.1097/PRS.0b013e318219c171] [Medline: 21701348]
2. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017 Feb 27;7(2):e012545 [FREE Full text] [doi: 10.1136/bmjopen-2016-012545] [Medline: 28242767]
3. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun 2019 Dec;16:100443 [FREE Full text] [doi: 10.1016/j.conctc.2019.100443] [Medline: 31497675]
4. Créquit P, Trinquart L, Yavchitz A, Ravaud P. Wasted research when systematic reviews fail to provide a complete and up-to-date evidence synthesis: the example of lung cancer. BMC Med 2016 Jan 20;14:8 [FREE Full text] [doi: 10.1186/s12916-016-0555-0] [Medline: 26792360]
5. Wright RW, Brand RA, Dunn W, Spindler KP. How to write a systematic review. Clin Orthop Relat Res 2007 Feb;455:23-29. [doi: 10.1097/BLO.0b013e31802c9098] [Medline: 17279036]
6. Clark J, McFarlane C, Cleo G, Ishikawa Ramos C, Marshall S. The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: case study. JMIR Med Educ 2021 May 31;7(2):e24418 [FREE Full text] [doi: 10.2196/24418] [Medline: 34057072]
7. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev 2014 Jul 09;3:74 [FREE Full text] [doi: 10.1186/2046-4053-3-74] [Medline: 25005128]
8. Marshall C, Sutton A. The systematic review toolbox. SR Tool Box. URL: http://www.systematicreviewtools.com/ [accessed 2021-08-27]
9. van der Mierden S, Tsaioun K, Bleich A, Leenaars CH. Software tools for literature screening in systematic reviews in biomedical research. ALTEX 2019;36(3):508-517. [doi: 10.14573/altex.1902131] [Medline: 31113000]
10. Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA. Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol 2020 Jan 13;20(1):7 [FREE Full text] [doi: 10.1186/s12874-020-0897-3] [Medline: 31931747]
11. Marshall C, Brereton P, Kitchenham B. Tools to support systematic reviews in software engineering: a feature analysis. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. 2014 Presented at: EASE '14; May 13-14, 2014; London, UK p. 1-10. [doi: 10.1145/2601248.2601270]
12. Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, et al. Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on CADIMA and review of existing tools. Environ Evid 2018 Feb 1;7(1):8. [doi: 10.1186/s13750-018-0115-5]
13. Nested knowledge. URL: https://nested-knowledge.com/ [accessed 2021-08-23]
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
15. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 2019 Jul 11;8(1):163 [FREE Full text] [doi: 10.1186/s13643-019-1074-9] [Medline: 31296265]
16. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol 2017 Nov 28;17(1):152 [FREE Full text] [doi: 10.1186/s12874-017-0431-4] [Medline: 29179685]
17. Vandvik PO, Brignardello-Petersen R, Guyatt GH. Living cumulative network meta-analysis to reduce waste in research: a paradigmatic shift for systematic reviews? BMC Med 2016 Mar 29;14:59 [FREE Full text] [doi: 10.1186/s12916-016-0596-4] [Medline: 27025849]
18. Bakken S. The journey to transparency, reproducibility, and replicability. J Am Med Inform Assoc 2019 Mar 01;26(3):185-187 [FREE Full text] [doi: 10.1093/jamia/ocz007] [Medline: 30689885]
19. Reed DD, Azulay RL. A microsoft excel(®) 2010 based tool for calculating interobserver agreement. Behav Anal Pract 2011;4(2):45-52 [FREE Full text] [doi: 10.1007/BF03391783] [Medline: 22649578]
20. Chai K, Ng L. Research Screener. URL: http://www.researchscreener.com/ [accessed 2021-08-27]

21. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev 2015 Jun 15;4:78 [FREE Full text] [doi: 10.1186/s13643-015-0066-7] [Medline: 26073888]
22. Pratt R, Gyllstrom B, Gearin K, Lange C, Hahn D, Baldwin LM, et al. Identifying barriers to collaboration between primary care and public health: experiences at the local level. Public Health Rep 2018;133(3):311-317 [FREE Full text] [doi: 10.1177/0033354918764391] [Medline: 29614236]
23. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, Living Systematic Review Network. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol 2017 Nov;91:23-30. [doi: 10.1016/j.jclinepi.2017.08.010] [Medline: 28912002]
24. Mavergames C, Elliott J. Living systematic reviews: towards real-time evidence for health-care decision-making. BMJ Publishing Group Limited. URL: https://bestpractice.bmj.com/info/us/toolkit/discuss-ebm/living-systematic-reviews-towards-real-time-evidence-for-health-care-decision-making/ [accessed 2021-08-27]
25. Kitchenham B, Linkman S, Law D. DESMET: a methodology for evaluating software engineering methods and tools. Comput Control Eng J 1997;8(3):120-126. [doi: 10.1049/cce:19970304]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SR: systematic review

Edited by C Lovis; submitted 28.08.21; peer-reviewed by N Eslamiamirabadi, A Brown; comments to author 14.11.21; revised version received 06.01.22; accepted 12.03.22; published 02.05.22.

Please cite as:
Cowie K, Rahmatullah A, Hardy N, Holub K, Kallmes K
Web-Based Software Tools for Systematic Literature Review in Medicine: Systematic Search and Feature Analysis
JMIR Med Inform 2022;10(5):e33219
URL: https://medinform.jmir.org/2022/5/e33219
doi:10.2196/33219
PMID:35499859

Original Paper

Impact of a Machine Learning–Based Decision Support System for Urinary Tract Infections: Prospective Observational Study in 36 Primary Care Practices

Willem Ernst Herter[1,2], BSc; Janine Khuc[2], MSc; Giovanni Cinà[2], MSc, PhD; Bart J Knottnerus[3], MSc, MD, PhD; Mattijs E Numans[1], MSc, MD, PhD, Prof Dr; Maryse A Wiewel[1,2], MSc, MD, PhD; Tobias N Bonten[1], MSc, MD, PhD; Daan P de Bruin[2], MSc; Thamar van Esch[3], MSc, PhD; Niels H Chavannes[1], MSc, MD, PhD, Prof Dr; Robert A Verheij[3], PhD, Prof Dr

[1]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, Netherlands

[2]Pacmed, Amsterdam, Netherlands

[3]Nivel Netherlands Institute for Health Services Research, Utrecht, Netherlands

Corresponding Author:
Willem Ernst Herter, BSc
Department of Public Health and Primary Care
Leiden University Medical Center
Albinusdreef 2
Leiden, 2333 ZA
Netherlands
Phone: 31 629292797
Email: w.e.herter@lumc.nl

Abstract

Background: There is increasing attention on machine learning (ML)-based clinical decision support systems (CDSS), but their added value and pitfalls are very rarely evaluated in clinical practice. We implemented a CDSS to aid general practitioners (GPs) in treating patients with urinary tract infections (UTIs), which are a significant health burden worldwide.

Objective: This study aims to prospectively assess the impact of this CDSS on treatment success and change in antibiotic prescription behavior of the physician. In doing so, we hope to identify drivers and obstacles that positively impact the quality of health care practice with ML.

Methods: The CDSS was developed by Pacmed, Nivel, and Leiden University Medical Center (LUMC). The CDSS presents the expected outcomes of treatments, using interpretable decision trees as ML classifiers. Treatment success was defined as a subsequent period of 28 days during which no new antibiotic treatment for UTI was needed. In this prospective observational study, 36 primary care practices used the software for 4 months. Furthermore, 29 control practices were identified using propensity score-matching. All analyses were performed using electronic health records from the Nivel Primary Care Database. Patients for whom the software was used were identified in the Nivel database by sequential matching using CDSS use data. We compared the proportion of successful treatments before and during the study within the treatment arm. The same analysis was performed for the control practices and the patient subgroup the software was definitely used for. All analyses, including that of physicians' prescription behavior, were statistically tested using 2-sided *z* tests with an α level of .05.

Results: In the treatment practices, 4998 observations were included before and 3422 observations (of 2423 unique patients) were included during the implementation period. In the control practices, 5044 observations were included before and 3360 observations were included during the implementation period. The proportion of successful treatments increased significantly from 75% to 80% in treatment practices (z=5.47, P<.001). No significant difference was detected in control practices (76% before and 76% during the pilot, z=0.02; P=.98). Of the 2423 patients, we identified 734 (30.29%) in the CDSS use database in the Nivel database. For these patients, the proportion of successful treatments during the study was 83%—a statistically significant difference, with 75% of successful treatments before the study in the treatment practices (z=4.95; P<.001).

Conclusions: The introduction of the CDSS as an intervention in the 36 treatment practices was associated with a statistically significant improvement in treatment success. We excluded temporal effects and validated the results with the subgroup analysis

in patients for whom we were certain that the software was used. This study shows important strengths and points of attention for the development and implementation of an ML-based CDSS in clinical practice.

Trial Registration: ClinicalTrials.gov NCT04408976; https://clinicaltrials.gov/ct2/show/NCT04408976

(JMIR Med Inform 2022;10(5):e27795) doi:10.2196/27795

KEYWORDS

machine learning; ML; artificial intelligence; clinical decision support system; implementation study; information technology; urinary tract infections

Introduction

Background

The application of machine learning (ML) in health care is increasing. Previous studies have shown that using data from electronic health records (EHRs) can inform us about treatment effectiveness and outcomes in a real patient population, providing insight into unknown disease correlations in the process [1-3]. As current medical knowledge is often based on average results from studies in an isolated clinical setting, these data could fill important knowledge gaps in practice resulting from the fact that randomized controlled trials often use stringent selection criteria and therefore do not cover the complexity and variety of patients in everyday practice [4-9].

Most algorithms featured in academic research do not reach clinical practice nor are their performances evaluated prospectively [10-12]. This makes it challenging to assess the true added value of ML in health care as well as to formulate a scientific and societal vision on the balance between this added value and its pitfalls and risks. Finally, little research has been conducted on the interaction of a clinical decision support system (CDSS) with the end user, which greatly affects adoption and clinical results [13-16].

The treatment of urinary tract infections (UTIs) in primary care offers an opportunity to add clinical value to ML. UTIs are common and represent a significant health burden worldwide [17,18]. In the Netherlands, a UTI is the most frequent diagnosis in women consulting general practitioners (GPs), with an incidence rate of 125 per 1000 patient years and 19.6 per 1000 patient years for men in 2018 [19]. Uncomplicated UTIs often occur in young, healthy, and nonpregnant women. Certain host factors predispose to the development of a complicated course, including abnormalities of the urinary tract, male sex, diabetes mellitus, immune deficiency, or immune-compromising drugs [18,20,21]. The treatment guidelines for patients with UTIs were published by the Dutch College of General Practitioners (NHG). At the time of this research, guidelines published in 2013 were in place [22]. Most clinical trials on the treatment of UTIs that underpin the evidence in this guideline are conducted on female patients with uncomplicated infections; hence, the scientific evidence for clinically effective treatments with increased risk of complicated UTIs is limited [20-22]. GPs consider the lack of agreement as a problem for all key recommendations while using UTI guidelines [23].

The development of ML-based algorithms could facilitate better decision-making through the delivery of individualized recommendations based on real-world data on all types of patients, which could be beneficial in determining the optimal treatment for patients at risk for complicated UTIs [11].

Supporting GPs With ML

Pacmed, a Dutch organization developing and implementing ML-based decision support in health care, developed, together with the consortium that conducted this research, a CDSS to aid GPs with the treatment choice for patients with a UTI. On the basis of the EHR data from UTI observations in the Nivel Primary Care Database, ML-based classifiers were constructed to estimate the probability of success of the 8 antibiotics commonly used for an individual patient with a UTI.

Study Objective

In this study, we prospectively assessed the impact of the CDSS on the clinical results and prescription behavior of physicians. For this purpose, we compared the proportion of successful treatments before and during the implementation of the CDSS as well as the proportion of antibiotics chosen by the physician. By conducting an implementation study among GPs in 36 practices in the Netherlands, we aim not only to assess the impact of the software but also to study the interaction and adoption of the software. In doing so, we hope to identify general drivers and obstacles that positively impact health care with ML.

Methods

Study Design

This research was carried out following a routine practice-based prospective observational study design, in which 36 practices used the software (henceforth, the treatment practices) for a period of 4 months, starting in November 2017. A period of 4 months was chosen based on a power analysis of the primary outcome as well as the prevalence of patients with UTI in Dutch primary care. Treatment practices were mostly recruited at the care group level. This is a partnership between primary care practices to collaboratively organize care for chronic diseases. These groups also often decide to collectively participate in innovation projects such as this research, without consulting every individual GP or primary care practice. Physicians from all participating practices were trained on the responsible use of the software and were instructed to its intended use as supportive to their decisions (ClinicalTrials.gov NCT04408976).

Ethics Approval

The study protocol was reviewed and determined to meet the requirements for exemption from the Ethics Committee (the Medical Ethical Committee) review under the Dutch Medical

Research Involving Human Subjects Act (WMO) and to be in accordance with the Dutch Medical Treatment Act (WGBO) and the Dutch Data Protection Act (WBP, now AVG).

The Clinical Decision Support Software: ML-Based Classifiers

The decision support system was developed through iterative consultation with multiple clinical stakeholders. A complete description of the model development and evaluation process is beyond the scope of this study. However, we provide some background information in the following paragraphs, highlighting the envisioned interaction with the end user in practice.

On the basis of the EHR data of patients who had at least one UTI between 2012 and 2014 and were >12 years, ML classifiers were constructed to estimate the probability of success for the 8 antibiotics commonly used for an individual patient with a UTI. In the potential absence of reliable UTI diagnosis data, we selected only patients who received antibiotic treatment for a UTI as reliably diagnosed patients in the data. Successful treatment was defined as a subsequent period of 28 days in which no new treatment was needed. The final data set for CDSS development contained 122,203 UTIs pertaining to 264 practices.

Owing to the anatomical differences between male and female patients with UTI, separate models were constructed for each sex. Fosfomycin was excluded as a treatment option in the clinical decision support system for male patients as this treatment is almost never used for male patients. The prediction would thus be of limited relevance, and the data set lacked sufficient data points to train a model. This approach resulted in 15 models in total: 8 for female patients and 7 for male patients.

The information presented by the ML classifiers was to be used in synergy with the existing experience and all other relevant sources of information. Hence, interpretability, clinical readability, and clinical relevance were prioritized in the development of the ML models. Decision trees were chosen as the classification method to allow for nonlinearities in the model while retaining interpretability. All 67 features that had been added as features to the classifiers were deemed medically important by the NHG guidelines issued at that time or had been indicated to affect treatment decisions, as discussed with medical experts [22]. These variables include patient characteristics, such as the presence of diabetes, pregnancy, indications of tissue invasion, dysfunctional urinary tracts, medical UTI history, and the treatment associated with these episodes. Other predictive features that were more difficult to interpret medically were also excluded. The classifiers were constructed using a scikit-learn pipeline, including missing value imputation, feature scaling, and L1 feature selection [24]. Hyperparameters were optimized using 10-fold cross-validation, and the model performance was evaluated using a cross-validated area under the receiver operator curve. This approach yielded modest model performance in terms of area under the curve (averaging around 0.6 over all models).

Although the classifiers were not able to predict with high accuracy which treatments would certainly (not) be successful, the models allowed for distinguishing patients with a relatively high risk of unsuccessful outcomes from patients with a low risk of unsuccessful treatment. More importantly, for a single patient, the models distinguished between treatments with a relatively high risk of unsuccessful outcomes and treatments with a medium or low risk of unsuccessful outcomes. Thus, although a substantial part of the outcome variation is unexplained, we expected the use of the model's predictions for treatment decisions to positively affect treatment outcomes.

The medical soundness and relevance of the decision trees were confirmed by multiple clinical experts inspecting the features and resulting models through a long list of clinical hypotheses on the practical performance of treatments for different patient groups. Moreover, before the implementation study, a (the Medical Ethical Committee') passive model validation was performed. Showing the predictions as well as the support information of the models for patients with UTI treated less than a month ago by their physician, we validated the usability, relevance, reliability, and interpretability of the information presented. These rounds of validations with medical experts convinced us that the models were reliable and could add value to clinical practice.

The Clinical Decision Support Software: User Interface

The software interface was developed in close collaboration with several primary care physicians through user tests and expert groups. The CDSS was not integrated into the EHR of GPs, so users were requested to enter patient characteristics into the web-based software. To invite the end user to interpret the information thoroughly and in an unbiased manner, antibiotics were always presented in the same order, independent of the probability of treatment success. Users were provided with a bar chart showing the estimated outcomes for the relevant treatment options based on similar patients within the database (Figure 1).

The algorithms in the CDSS presented the expected outcomes as well as the necessary support information for the physician at the time of the treatment decision. Owing to the choice of decision tree classifiers, we were able to follow the characteristics of an individual patient through the decision tree nodes and share with the physician the characteristics of the sample which was used to predict the outcome of a treatment. This information, such as the age range of these patients and other clinically relevant features, can be retrieved by clicking on the relevant treatment.

We chose not to display the models themselves because the large number of features and the different models would have been confusing. We did not add CIs around the predictions in the user interface. Calculating and presenting a CI around a probabilistic prediction is not straightforward, and this complexity could have been confusing or incorrectly interpreted.

In addition to the presentation of the expected outcomes, the relevant part of the 2013 NHG guidelines was also presented. All 8 antibiotics from the NHG recommendations for patients with a UTI are shown, although for female patients without

signs of tissue invasion, after expert consultation, it was decided not to show antibiotics with high tissue penetration as treatment options, as other treatments should be considered and tried first in most instances. Finally, GPs were instructed to use the software only for patients >12 years and to assess the information presented together with all other information they deemed relevant in treating patients with a UTI.

Figure 1. Decision support software: interface to enter patient characteristics (top); presentation of expected outcomes and NHG (Dutch College of General Practitioners) guidelines (bottom).

Selection of Control Practices

Control practices were identified from a pool of 129 potential control practices in the Nivel database through a propensity score-matched augmented control procedure. As shown in previous research, these matching methods can be used to construct an artificial control group for trials by matching treatment and control units that are similar in terms of their observable characteristics [25,26]. As practice characteristics are most informative in the way patients are being treated, propensity score-matching was performed at the aggregated practice level. The total number of patients per practice and their average age were the characteristics used to construct the propensity. Matching was performed using a caliper of 0.05.

The treatment practices were matched with 29 control practices. Data from all patients with a UTI in these practices were analyzed. This resulted in 4998 observations of patients with a UTI in the treatment practices before the pilot started and 3422 observations during the pilot. The control practices resulted in 5044 observations before and 3360 observations during the pilot.

Preparation of Study Data

Nivel Primary Care Database

All analyses were performed using the Nivel Primary Care Database, containing structured EHR data from 530 GP practices in the Netherlands. A selection was made so that the data set only contained the data of all patients with at least one UTI during this study.

The data request was approved by all necessary and appropriate bodies from the Nivel Governance-Document Nivel Primary Care Database [27] under number NZR00317.030. The use of the data for this specific research was in accordance with all relevant Dutch and European laws and legislations.

The study included all patients who had at least one indication of UTI symptoms, indicated by the International Classification of Primary Care codes U70 (acute pyelonephritis), U71 (cystitis), U72 (nonspecific urethritis), U01 (painful micturition), or U02 (frequent micturition) and were prescribed antibiotic treatment for this UTI.

We used data from 2 periods. The first period consisted of the time before the implementation study (week 16 until week 4), and the second period consisted of the time during the implementation period (weeks 0 to 20). Within these 20 weeks, the software was used for different periods of 16 weeks. Owing to the defined outcome measure of treatment success that requires patients not to receive another treatment for UTI within 28 days, an additional 4 weeks were added to ensure that the treatment outcome of these patients was also captured within the analysis.

The prescribed treatment for UTI was directly recorded in the EHR system of the GP. Background information about the patient and their comorbidities consisted of a combination of diagnoses and symptom codes and the prescription of other medications related to these comorbidities.

Pacmed Use Database

Through use of physicians and their assistants, data on patient characteristics and chosen treatments were generated using Pacmed software. These data were generated with informed consent from the treated patients.

The patients present in this database were those for whom we were certain that the software had been used. Therefore, we attempted to identify these *Pacmed patients* in the Nivel database. However, because all identifiable personal data were removed for both the Pacmed software and the Nivel database, it was not possible to match these 2 databases directly.

Instead, identification took place iteratively using a sequential matching procedure. A single matching approach failed because of practical challenges resulting from the nature of the data sets. The Nivel database was generated automatically from all the different information systems used in the participating practices, resulting in a data set with subtle differences between the practices. The Pacmed data consist of data directly resulting from the use of the CDSS. Attempting to match the databases through a single matching procedure based on practice location, gender, date of birth, and the data of consultation failed, as we assessed it as very likely that the Pacmed software had been used days after the first visit or at a second visit. Therefore, another approach was used where patient identification was performed iteratively through a sequential matching procedure, in such a way that after an initial merge, unique matches that were found were removed and the matching with a new variable constellation for the remaining patients was continued.

First, an effort was made to identify patients from Pacmed in the raw EHR data provided by Nivel. Thereafter, additional matching was performed for patients in the processed Nivel data set. Variables that were matched based on the raw Nivel data included age, sex, date of birth, and date of consultation. In addition, age categories, day intervals around the date of consultation, prescribed medication, and comorbidities were constructed. All matches included gender and practice postal codes.

Primary and Secondary Outcomes

The primary outcome of interest was the difference in the proportion of successful treatments between the periods before and during the study. Successful treatment was defined as a period of 28 days after the initial treatment, during which no new treatment was needed. This outcome definition was constructed using GP expert groups and a meticulous analysis of the impact of different definition choices.

The analyses of the primary outcomes were repeated for subgroups based on sex and age, if the patients had diabetes and if UTI was complicated. In addition, to directly compare the differences between the treatment and control arms, the primary outcome was compared between the groups during the implementation study.

A total of 2 sensitivity analyses were performed to test the robustness of primary outcomes. First, to exclude potential temporal effects, the same test for the primary outcome of interest was performed for control practices. Second, the proportion of successful treatments for patients for whom the software had certainly been used was compared with the proportion of successful treatments for patients in the treatment practices before the implementation study.

Finally, the prescription behavior of the physicians was analyzed to determine whether there was a statistically significant difference in prescribed antibiotics between the treatment and control practices before and during the implementation study period. This analysis was repeated for observations for which we were sure that the software had been used. Specifically, we were interested in the difference in the proportion of high tissue penetration antibiotics chosen by physicians when presented with the expected outcomes of treatments.

Statistical Analysis

Power Analysis

We conducted a power analysis to gauge the required number of observations of patients with UTIs to detect small effect sizes (Cohen effect size of 0.1) of the intervention [28]. Patients were clustered within a practice, making them more likely to be treated or respond similarly within that practice. To account for this clustering, sample sizes were adjusted using an inflation factor [25]. The inflation factor is a function of the intracluster correlation coefficient and average cluster sizes per practice. However, the mean average cluster size (the number of patients with a UTI) per practice was 72.7, and practices showed large variations in average cluster sizes (SD 98.3, range 10-325). Therefore, along with the inflation factor, a cluster variation coefficient was calculated for the outcome variables. Thus, the reported sample size was adjusted, including the intracluster correlation and cluster variation coefficients. The desired sample size was calculated to be at least 851 at a power of 0.8 and a type 1 error rate of 0.05 to detect small effects.

Outcome Statistics

A total of 2 sample *z* tests with an α level of .05 were used to test the statistical significance of the primary outcome analysis, both the sensitivity analyses and the prescription behavior analyses. To test the significance of the subgroup analyses for the primary outcome measure, additional *z* tests were used. To determine whether the differences were statistically significant and to avoid the inflation of type 1 errors, Bonferroni corrections were applied. Differences with a *P* value <.006 (0.05/9) were determined to be statistically significant. To further compare the primary outcomes across different subgroups, the relative risk ratio was used [29].

Results

Patient Population

The Pacmed use database contained 1689 unique patients, of which 734 (43.46%) unique individual patients were identified in the Nivel database through the sequential identification procedure. This is the number of patients for whom we can be certain that the software was used.

Figure 2 shows the variables used in the sequential matching procedure and the number of matches found in each iteration. Table 1 displays the patient characteristics for the cohort in treatment and control practices during the implementation study and specifies the characteristics of the patients identified from the Pacmed use database.

Figure 2. The number of matches found through the sequential matching procedure.

Iteration	Variables						Matches (n)
	Date of birth	Age	Date	Date interval	Treatment	Comorbidity	
Raw data							
1	■		■				277
2	■			■			47
3		■	■				176
4		■		■			68
Processed data							
5	■		■		■		12
6	■		■			■	1
7	■			■		■	0
8	■			■	■		2
9		■	■		■		8
10		■	■			■	8
11		■		■	■		2
12		■		■		■	0
13			■			■	78
14				■		■	55
Total							734

Table 1. Patient characteristics of the observations in treatment and control practices during the implementation study.

Characteristic	Proportion of patients in treatment practices observations (n=3422)	Proportion of patients identified in Pacmed Clinical Decision Support System observations (n=1121)	Proportion of patients in control practices observations (n=3360)
Sex (female)	0.88	0.92	0.86
Age (years)			
<30	0.17	0.15	0.17
30-50	0.21	0.18	0.20
50-70	0.36	0.39	0.29
>70	0.26	0.28	0.33
Diabetes	0.08	0.09	0.13
UTI[a] with tissue invasion[b]	0.11	0.13	0.12
Complicated UTI[c]	0.12	0.13	0.13

[a]UTI: urinary tract infection.

[b]A UTI with tissue invasion was defined as a UTI with which (a combination of) the International Classification of Primary Care codes associated with tissue invasion-related symptoms were registered (A02, A03, A04, and A05).

[c]Complicated UTI is defined as a UTI with tissue invasion or a simultaneous pyelonephritis or prostatitis episode, International Classification of Primary Care codes U70 and Y93, respectively.

Evaluation Outcomes

Primary Outcome and Sensitivity Analyses

The proportion of successful treatments increased significantly from 75% to 80% in treatment practices (z=5.47; P<.001). In the control practices, no significant change in outcomes was observed during the same period (76% before and 76% during the pilot, z=0.02; P=.98). The proportion of successful treatments during the study was 83% for the observations of which we are certain that the software had been used. This was a statistically significant difference, with 75% of successful treatments before the study in the treatment practices (z=4.95; P<.001). The comparison of the primary outcome between the control practices and the treatment practices during the implementation study also showed a significant difference (76% for the treatment practices and 80% for the control practices, z=4.86; P<.001).

The change in outcome has been specified for subgroups based on sex, age, comorbidities (diabetes) and whether the UTI was complicated in Table 2. In this analysis, the increase in outcomes was statistically significant for female patients and patients >70 years.

Table 2. Test statistics of primary outcome or several patient subgroups in the treatment practices observations before (n=4998) and during (n=3422) the study.

Subgroup	Proportion of successful treatments before study	Proportion of successful treatments during study	Risk ratio	*P* value
Sex				
Female (n=3008)	0.76	0.81	1.07	<.001
Male (n=414)	0.72	0.78	1.08	.01
Age (years)				
<30 (n=580)	0.81	0.86	1.06	.01
30-50 (n=719)	0.80	0.83	1.04	.05
50-70 (n=1227)	0.76	0.79	1.04	.05
>70 (n=896)	0.70	0.76	1.09	<.001[a]
Complicated UTI[b,c] (n=404)	0.72	0.79	1.10	.03
Diabetes (n=275)	0.71	0.79	1.11	.03

[a]Significant Bonferroni adjusted *P* values (.05/9).

[b]UTI: urinary tract infection.

[c]A complicated UTI is defined as a UTI with tissue invasion or a simultaneous pyelonephritis or prostatitis episode, International Classification of Primary Codes U70 and Y93, respectively.

GP Prescription Behavior

In the treatment practices as well as in the control practices, there was no significant difference in the proportion of high tissue penetration antibiotics prescribed between the period prior and during the implementation study. Table 3 has additional information on the choice of treatment before and during the study period. As it is known that there are sex differences in prescribed medications owing to differences in underlying etiology, the same table is shown for both sexes. Table 4 shows the same information for the observations for which we were sure that the software was used.

Table 3. Proportion of medication prescribed before and during implementation study.

	Treatment practices				Control practices			
	Before	During	Delta %	*P* value	Before	During	Delta %	*P* value
All patients	n=4998	n=3422			n=5044	n=3360		
Antibiotics with low tissue penetration	0.79	0.80	0.01	.37	0.79	0.77	−0.02	.02
Nitrofurantoin	0.59	0.59	−0.01	.56	0.60	0.57	−0.02	.03
Fosfomycin	0.13	0.13	0.00	.70	0.14	0.15	0.01	.24
Trimethoprim	0.06	0.06	0.00	.58	0.05	0.04	−0.01	.11
Norfloxacin	0.01	0.02	0.02	.001	0.00	0.00	0.00	.63
Antibiotics with high tissue penetration	0.21	0.20	−0.01	.37	0.21	0.23	0.02	.02
Ciprofloxacin	0.16	0.14	−0.02	.04	0.13	0.15	0.01	.09
Augmentin	0.03	0.03	0.00	.44	0.04	0.04	0.00	.59
Sulfamethoxazole and trimethoprim	0.02	0.02	0.00	.65	0.02	0.03	0.01	.005
Amoxicillin	0.01	0.01	0.00	.45	0.01	0.01	0.00	.58
Female patients	n=4361	n=3008			n=4439	n=2881		
Antibiotics with low tissue penetration	0.84	0.84	0.00	.99	0.84	0.83	−0.01	.12
Nitrofurantoin	0.62	0.61	−0.01	.39	0.63	0.62	−0.02	.11
Fosfomycin	0.14	0.14	0.00	.75	0.15	0.16	0.01	.12
Trimethoprim	0.06	0.06	0.00	.99	0.05	0.04	−0.01	.09
Norfloxacin	0.01	0.01	0.01	.006	0.00	0.00	0.00	.69
Antibiotics with high tissue penetration	0.16	0.16	0.00	.99	0.16	0.17	0.01	.12
Ciprofloxacin	0.12	0.11	0.00	.66	0.10	0.10	0.00	.61
Augmentin	0.02	0.02	0.00	.55	0.10	0.10	0.00	.61
Sulfamethoxazole and trimethoprim	0.01	0.01	0.00	.98	0.02	0.02	0.01	.01
Amoxicillin	0.01	0.01	0.00	.77	0.01	0.01	0.00	.39
Male patients	n=637	n=414			n=605	n=478		
Antibiotics with low tissue penetration	0.44	0.49	0.05	.14	0.44	0.43	−0.01	.77
Nitrofurantoin	0.36	0.37	0.01	.86	0.33	0.33	−0.01	.84
Fosfomycin	0.03	0.03	0.00	.94	0.06	0.06	−0.00	.92
Trimethoprim	0.03	0.05	0.02	.08	0.04	0.04	0.00	.97
Norfloxacin	0.02	0.04	0.02	.09	0.01	0.01	0.00	.70
Antibiotics with high tissue penetration	0.56	0.51	−0.05	.14	0.56	0.57	0.01	.77
Ciprofloxacin	0.44	0.35	−0.09	.003	0.39	0.42	0.02	.41
Augmentin	0.07	0.08	0.01	.52	0.09	0.06	−0.02	.13
Sulfamethoxazole and trimethoprim	0.03	0.05	0.01	.37	0.06	0.07	0.01	.36
Amoxicillin	0.01	0.02	0.02	.05	0.02	0.01	−0.01	.50
Patients with complicated UTI[a,b]	n=365	n=404			n=369	n=433		
Antibiotics with low tissue penetration	0.58	0.62	0.05	.19	0.62	0.61	−0.01	.74
Nitrofurantoin	0.42	0.43	0.01	.80	0.43	0.36	−0.06	.07
Fosfomycin	0.10	0.10	0.00	.91	0.10	0.21	0.12	.92
Trimethoprim	0.05	0.06	0.02	.37	0.06	0.03	−0.04	.02
Norfloxacin	0.01	0.03	0.02	.01	0.03	0.00	−0.03	.003
Antibiotics with high tissue penetration	0.42	0.38	−0.05	.19	0.38	0.39	0.01	.74
Ciprofloxacin	0.30	0.22	−0.08	.01	0.22	0.22	0.00	.97

	Treatment practices				Control practices			
	Before	During	Delta %	*P* value	Before	During	Delta %	*P* value
Augmentin	0.07	0.08	0.01	.77	0.08	0.10	0.02	.26
Sulfamethoxazole and trimethoprim	0.03	0.05	0.02	.11	0.05	0.05	0.00	.93
Amoxicillin	0.02	0.02	0.00	.76	0.02	0.02	−0.01	.53
Patients with diabetes	n=355	n=275			n=359	n=429		
Antibiotics with low tissue penetration	0.70	0.72	0.01	.74	0.75	0.68	−0.07	.02
Nitrofurantoin	0.47	0.41	−0.06	.14	0.52	0.43	−0.09	.01
Fosfomycin	0.14	0.14	0.00	.92	0.17	0.18	0.01	.65
Trimethoprim	0.06	0.10	0.04	.10	0.06	0.06	0.01	.67
Norfloxacin	0.03	0.06	0.04	.03	0.01	0.01	0.00	.83
Antibiotics with high tissue penetration	0.30	0.28	−0.01	.74	0.25	0.32	0.07	.02
Ciprofloxacin	0.25	0.22	−0.03	.43	0.17	0.21	0.04	.16
Augmentin	0.03	0.04	0.00	.86	0.04	0.04	0.00	.88
Sulfamethoxazole and trimethoprim	0.01	0.02	0.00	.69	0.02	0.05	0.03	.03
Amoxicillin	0.00	0.01	0.01	.08	0.01	0.01	0.01	.45
Patients with age >70 years	n=1599	n=896			n=1721	n=1122		
Antibiotics with low tissue penetration	0.70	0.72	0.02	.31	0.74	0.70	−0.04	.02
Nitrofurantoin	0.45	0.43	−0.03	.20	0.48	0.45	−0.03	.11
Fosfomycin	0.16	0.17	0.01	.48	0.20	0.19	−0.01	.74
Trimethoprim	0.07	0.07	0.01	.47	0.06	0.06	−0.01	.47
Norfloxacin	0.02	0.05	0.03	.001	0.00	0.01	0.00	.58
Antibiotics with high tissue penetration	0.30	0.28	−0.02	.31	0.26	0.30	0.04	.02
Ciprofloxacin	0.24	0.21	−0.03	.06	0.17	0.20	0.03	.03
Augmentin	0.03	0.04	0.01	.39	0.05	0.03	−0.02	.03
Sulfamethoxazole and trimethoprim	0.02	0.03	0.01	.32	0.02	0.04	0.02	.004
Amoxicillin	0.01	0.01	0.00	.75	0.01	0.02	0.01	.27

[a]UTI: urinary tract infection.

[b]A complicated UTI is defined as a UTI with tissue invasion or a simultaneous pyelonephritis or prostatitis episode, International Classification of Primary Care codes U70 and Y93, respectively.

Table 4. Proportion of medication prescribed before and during implementation study for all patients, for the observations identified from the Pacmed use database.

	Treatment practices			
	Before (n=4998)	During (identified; n=1121)	Delta %	*P* value
Antibiotics with low tissue penetration	0.79	0.81	0.02	.15
Nitrofurantoin	0.59	0.53	−0.07	<.001
Fosfomycin	0.13	0.16	0.03	<.001
Trimethoprim	0.06	0.08	0.02	.01
Norfloxacin	0.01	0.04	0.03	<.001
Antibiotics with high tissue penetration	0.21	0.19	−0.02	.15
Ciprofloxacin	0.16	0.13	−0.03	.01
Augmentin	0.03	0.03	0.00	.53
Sulfamethoxazole and trimethoprim	0.02	0.02	0.00	.39
Amoxicillin	0.01	0.01	0.01	.02

Discussion

Principal Findings

The most important result of this study is that the introduction of the CDSS as an intervention in the 34 treatment practices was associated with improved treatment success for patients with UTI. The percentage of successful treatments in the patient population increased from 75% before implementation of the CDSS to 80% during the implementation period. Next to a significant increase in treatment outcome within the treatment arm, the difference between treatment and control group was also significant (76% in control practices and 80% in treatment practices during the study). These control practices were selected through propensity score-matching and were similar at baseline. This resulted in control practices that, at the time of the study, had comparable patient populations with UTI.

To assess whether the increase in treatment success was due to the CDSS presented in this study, we performed 2 sensitivity analyses. First, the software was not used for all patients in the Nivel database. Therefore, we sought to identify patients from the Pacmed use database in the Nivel database. We had been able to identify more than half of the patients in the Pacmed use database in the Nivel database. The association of the increase in treatment success with the introduction of the intervention was strengthened by the fact that the increase in clinical outcome was even higher and statistically significant (83%) for the patients for whom we were sure that the software had been used. The significance of this test must be seen from the perspective that the populations are not identical, and that there is a potential selection bias in the patient population in the CDSS database. However, Table 1 shows comparable prevalence among the relevant clinical subgroups.

Second, we assessed whether the increase in treatment success was due to temporal effects, namely, the spontaneous improvement of treatment success for all practices, independent of the introduction of the CDSS. In the control practices, no significant increase in the proportion of successful treatments was observed.

Finally, the increase in treatment success did not seem to have been caused by an increase in the prescription of high tissue penetration antibiotics. On an average, for female and male patients, we did not observe a significant difference in the proportion of antibiotics with high tissue penetration. We observed a significant increase in the prescription of norfloxacin in female patients.

Behavior Change Within the Treatment Practices

An increase in treatment success was observed for multiple subgroups. In this analysis, only the results for female patients and patients aged >70 years were found to be statistically significant. Other subgroups with a noteworthy increase in outcomes included male patients, patients with complicated UTIs, and patients with diabetes. The reason that the increase in outcome cannot be deemed significant in this analysis is presumably partly owing to the sample sizes of these subgroups, as opposed to lower effect sizes. In particular, most clinical trials on the treatment of UTIs, supporting the evidence in the Dutch GP guidelines at the time, were conducted in female patients with uncomplicated infections [20-22]. One could then expect that an ML-based CDSS, aiming to fill knowledge gaps by learning from more complex patients in practice, would be most valuable for patient groups that are now understudied.

Within these subgroups, although not statistically significant, we observed an indication of behavioral change in the treatment arm. For all these subgroups (male, patients with diabetes, patients with a complicated UTI, and patients >70 years), the proportion of norfloxacin treatments doubled, which was not observed in the control practices. Norfloxacin was not recommended as a treatment option for all subgroups in the NHG guideline.

The NHG guidelines at the time recommended nitrofurantoin as the first choice for male and diabetic patients, with trimethoprim as the second choice. Trimethoprim prescriptions almost doubled for both subgroups only in the treatment practices. For patients with a complicated UTI, we observed a decrease in ciprofloxacin treatment only in the treatment practices, although ciprofloxacin was the first recommended treatment in the NHG guidelines at the time. Using Bonferroni adjusted *P* values, the difference in prescription behavior was not deemed to be statistically significant, and the effect of this indicated behavior change on clinical outcome should serve as a hypothesis for future research. However, it should be noted that it is unlikely that the increase in outcomes is (solely) owing to better guideline adherence. The analysis of behavioral changes for the CDSS patients identified in the Nivel database confirms this insight, with a significant decrease in nitrofurantoin treatments, which is the first recommended treatment option for almost all patient groups in the guideline, and a significant increase in norfloxacin treatments.

However, many other unmeasured factors could have improved patient outcomes independent of the information presented by the CDSS. Among other things, knowing to participate in the trial could have led to a better diagnosis and more conscious treatment choice, independent of the relevance or value of the CDSS.

Strengths

Only by integrating the knowledge in the clinician's workflow and evaluating the impact prospectively can we truly assess the potential added value of a new technology such as ML in today's health care system. A strength of this study is the fact that we developed a CDSS that, through its accessibility, was often used by the participating physicians to enable us to analyze the difference in our chosen outcome on a scale large enough for the results to be statistically significant. A more in-depth study on the use and perceived accessibility of physicians is in preparation.

In the design of the software, the transparency of the underlying technology was key to ensuring its usability. The algorithms that form the intelligence of the software were deliberately chosen to be interpretable and understandable models for the end user as well as for the physicians who were part of the development. In addition, the resulting predictions of treatment success presented in the CDSS during the study were

accompanied by supporting information regarding the patient characteristics on which the predictions were based. Finally, by presenting the relevant subsection of the active NHG guidelines, we presented the necessary context information to assess all sources of information equally, enabling the physician to combine these sources together with their experience, expertise, and intuition to make the best decision for the patient.

The execution of sensitivity analyses to assess the robustness of the association between treatment success and the intervention of the CDSS is another strength of this study. We had access to EHR from the Nivel database of treatment practices as well as from control practices, enabling us to exclude temporal effects. The fact that we had access to the Pacmed use database made it possible to specifically analyze the subgroup of patients for whom we were certain that the tool had been used.

We observed that for patients in whom the tool had been used, the outcome improvement seemed more pronounced compared with the rest of the patients in the treatment practices. This suggests a clear added value of the CDSS tool itself, rather than the mere fact that more attention was paid to UTI in these practices. Next, it is noteworthy that we were able to assess the adoption rate of the software and relate it to the outcome of interest. Only by doing so is it possible to evaluate the expected impact in relation to the (financial) investment needed to develop, implement, integrate, monitor, and continuously improve complex technologies such as ML for physicians [14,15].

Limitations

To assess the impact of the CDSS on treatment success, we chose several measurements to ensure the robustness and reliability of our outcome measure. To ensure that the patients included were indeed affected by UTI, we selected only patients on having received antibiotic treatment rather than using only the diagnosis code as the selection criterion. However, a more specific diagnosis of a UTI can be made through laboratory data, which have not been used, as laboratory test results were not recorded in the EHR data for many UTI patients. In addition, there was a selection bias in selecting only patients who received antibiotic treatment. However, this selection bias was the same for the treatment and control practices for this study.

Furthermore, treatment success is indirectly derived from the information systems of GPs and is used as a proxy for clinical examination. This method is similar to the algorithms used to construct disease episodes based on EHR data [30]. A similar methodology was followed, defining treatment success as a subsequent period of 28 days, where no new treatment was needed, indicating a reduced risk for treatment failure or relapse. Possible flaws concerning therapy compliance cannot be mitigated using these data. Information about the resolution of complaints or bacterial clearance, and thus, decisive knowledge on treatment success, was absent. In addition, it is likely that unnecessary antibiotic prescriptions were considered successful based on our definition of treatment success. However, the potential flaw in our outcome definition was consistent with treatment and control practices. Moreover, we have no indication that the use of the CDSS resulted in more unnecessary prescriptions (and with that positively influenced the primary outcome of this research). In both the treatment and control practices, the total number of antibiotic prescriptions decreased by almost identical proportions (from 4998 to 3422 in the treatment practices [–32%] and from 5044 to 3360 in the control practices [–33%]).

In addition, negative effects, side effects, or impacts on general resistance to treatments were not investigated in this study. These are, of course, factors that should impact the evaluation of a CDSS of this kind in practice. However, we excluded antibiotics with high tissue penetration in the software as treatment options, as we were aware of the potential negative effects of these treatments. This might have contributed to the fact that we did not observe a significant increase in these proportions of treatments prescribed during the study. Furthermore, it can be expected that the side effects of the treatments were considered by GPs in the treatment and control arms of the study. GPs were actively advised to choose a treatment based on all the information they deemed relevant, not only the information presented through the CDSS. Therefore, as we expect GPs to include knowledge on the side effects and negative effects of treatments in their decision-making, we mitigate the impact of this information being absent in the CDSS.

Finally, we were unable to relate the significant increase in patient outcomes to significant behavior change by physicians in treatment practices. The analyses of behavior change could potentially have been done more thoroughly if we had been able to match more patients between the Pacmed data and Nivel database. Out of 1200, only 734 (61.16%) patients could be matched owing to an underestimated complexity in matching these databases. In hindsight, a pseudonymized patient identifier would have enabled us to match a significantly higher number of patients in the Nivel database, if not all, from the CDSS data. The most important underestimation of this complexity was the incorrect assumption that participating physicians and assistants would always use the software during the first consultation with the patients. As this was not the case, it was difficult to identify patients based on their characteristics as well as the time they had entered into the software. We strongly recommend extensive research on the care paths of patients in multiple clinical institutions to match the implementation study design with these care paths. Another effort to further understand the behavior change and thus the impact of the CDSS would be to have expert groups and extensive surveys on how the software impacted the decision-making of GPs and assistants at an individual level.

Relevance and Future Directions

This research brings important knowledge to the research field of responsibly implemented ML-based decision support systems with clinical relevance. In particular, most published algorithms do not reach the frontline of clinical practice nor are they validated prospectively [10,11]. To make this technology live up to its great promises, prospective validation is needed, resulting in high-quality protocols for the responsible development and deployment of ML in health care [11,14,31-34].

Unfortunately, there is very little scientific discussion on the responsible evaluation of the CDSS to assess its impact and

validity once it has been implemented. Nevertheless, only by evaluating its impact on relevant outcomes, while integrated in the clinical workflow, the potential and risks can be fully understood [15,35]. One of the greatest barriers to achieving impact in practice is the adoption of the CDSS by the clinician as an end user [15,33]. Nevertheless, studies often fail to assess (or report on) the impact of the technology on the users and their workflow [13,14,16]. Therefore, it is even more important to have protocols in place to thoroughly analyze the behavior change, assess its robustness, and compare the outcomes for groups for which the tool has certainly been used with those that one is uncertain about.

Conclusions

The introduction of the CDSS as an intervention in the 36 treatment practices was associated with a statistically significant improvement in treatment success for patients with a UTI, namely, an increase from 75% to 80% successful treatments. The 2 sensitivity analyses enabled us to present this result with greater robustness. First, temporal effects were excluded by evaluating treatment success in the same period for a group of control practices selected through a propensity score-matching procedure. Second, analyzing the subgroup for whom we were certain the software had been used strengthened the association of an increase in successful treatment with the presentation of the CDSS in the treatment arm.

This study shows some important strengths and points of attention in the design and development of clinical decision support software as well as a thorough evaluation of its clinical impact in practice. Further research is needed on the interaction of ML-based clinical decision support software with end users to assess the potential impact of this technology for patients and physicians, and to develop concrete and objective guidelines to perform this research responsibly.

Acknowledgments

Zilveren Kruis, CZ Groep, and Menzis, 3 large health insurers in the Netherlands, partly funded this study. The funding sources had no role in the study design, analysis, interpretation of data, writing of the report, or submission of the article for publication.

Conflicts of Interest

WEH is a PhD candidate at Leiden University Medical Center (LUMC) as well as a director at Pacmed. JK, MAW, DPdB, and GC worked at Pacmed during their contributions to the research and paper.

References

1. 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;53:162-173 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
3. Krittanawong C. The rise of artificial intelligence and the uncertain future for physicians. Eur J Intern Med 2018;48:e13-e14. [doi: 10.1016/j.ejim.2017.06.017] [Medline: 28651747]
4. Durso SC. Using clinical guidelines designed for older adults with diabetes mellitus and complex health status. JAMA 2006;295(16):1935-1940. [doi: 10.1001/jama.295.16.1935] [Medline: 16639053]
5. Roche N, Anzueto A, Bosnic Anticevich S, Kaplan A, Miravitlles M, Ryan D, Respiratory Effectiveness Group Collaborators. The importance of real-life research in respiratory medicine: manifesto of the Respiratory Effectiveness Group: endorsed by the International Primary Care Respiratory Group and the World Allergy Organization. Eur Respir J 2019;54(3):1901511 [FREE Full text] [doi: 10.1183/13993003.01511-2019] [Medline: 31537655]
6. Shaneyfelt TM, Centor RM. Reassessment of clinical practice guidelines: go gently into that good night. JAMA 2009;301(8):868-869. [doi: 10.1001/jama.2009.225] [Medline: 19244197]
7. Tinetti ME, Bogardus Jr ST, Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. N Engl J Med 2004;351(27):2870-2874. [doi: 10.1056/NEJMsb042458] [Medline: 15625341]
8. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA 2005;294(6):716-724. [doi: 10.1001/jama.294.6.716] [Medline: 16091574]
9. Schoen C, Osborn R, Huynh PT, Doty M, Peugh J, Zapert K. On the front lines of care: primary care doctors' office systems, experiences, and views in seven countries. Health Aff (Millwood) 2006;25(6):w555-w571. [doi: 10.1377/hlthaff.25.w555] [Medline: 17102164]
10. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019;2:77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]
11. Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure FX, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 2020;26(5):584-595 [FREE Full text] [doi: 10.1016/j.cmi.2019.09.009] [Medline: 31539636]

12. Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, et al. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. BMJ 2013;346:f657 [FREE Full text] [doi: 10.1136/bmj.f657] [Medline: 23412440]
13. Rawson TM, Moore LS, Hernandez B, Charani E, Castro-Sanchez E, Herrero P, et al. A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately? Clin Microbiol Infect 2017;23(8):524-532 [FREE Full text] [doi: 10.1016/j.cmi.2017.02.028] [Medline: 28268133]
14. Sennesael AL, Krug B, Sneyers B, Spinewine A. Do computerized clinical decision support systems improve the prescribing of oral anticoagulants? A systematic review. Thromb Res 2020;187:79-87. [doi: 10.1016/j.thromres.2019.12.023] [Medline: 31972381]
15. Keyworth C, Hart J, Armitage CJ, Tully MP. What maximizes the effectiveness and implementation of technology-based interventions to support healthcare professional practice? A systematic literature review. BMC Med Inform Decis Mak 2018;18(1):93 [FREE Full text] [doi: 10.1186/s12911-018-0661-3] [Medline: 30404638]
16. Naqa IE, Kosorok MR, Jin J, Mierzwa M, Ten Haken RK. Prospects and challenges for clinical decision support in the era of big data. JCO Clin Cancer Inform 2018;2:1-12 [FREE Full text] [doi: 10.1200/CCI.18.00002] [Medline: 30613823]
17. Foxman B. The epidemiology of urinary tract infection. Nat Rev Urol 2010;7(12):653-660. [doi: 10.1038/nrurol.2010.190] [Medline: 21139641]
18. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. Nat Rev Microbiol 2015;13(5):269-284 [FREE Full text] [doi: 10.1038/nrmicro3432] [Medline: 25853778]
19. Jaarcijfers aandoeningen - Huisartsenregistraties. Nivel. URL: https://www.nivel.nl/nl/nivel-zorgregistraties-eerste-lijn/jaarcijfers-aandoeningen-huisartsenregistraties [accessed 2020-12-24]
20. Gupta K, Hooton TM, Naber KG, Wullt B, Colgan R, Miller LG, Infectious Diseases Society of America, European Society for Microbiology and Infectious Diseases. International clinical practice guidelines for the treatment of acute uncomplicated cystitis and pyelonephritis in women: a 2010 update by the Infectious Diseases Society of America and the European Society for Microbiology and Infectious Diseases. Clin Infect Dis 2011;52(5):e103-e120. [doi: 10.1093/cid/ciq257] [Medline: 21292654]
21. Schneeberger C, Stolk RP, Devries JH, Schneeberger PM, Herings RM, Geerlings SE. Differences in the pattern of antibiotic prescription profile and recurrence rate for possible urinary tract infections in women with and without diabetes. Diabetes Care 2008;31(7):1380-1385 [FREE Full text] [doi: 10.2337/dc07-2188] [Medline: 18362200]
22. van Pinxteren B, Knottnerus B, Geerlings S, Visser I, Klinkhamer S. NHG-Standaard Urineweginfecties (derde herziening). Nederlands Huisartsen Genootschap. 2013. URL: https://www.nhg.org/actueel/nieuws/herziening-nhg-standaard-urineweginfecties [accessed 2022-03-24]
23. Lugtenberg M, Zegers-van Schaick JM, Westert GP, Burgers JS. Why don't physicians adhere to guideline recommendations in practice? An analysis of barriers among Dutch general practitioners. Implement Sci 2009;4:54 [FREE Full text] [doi: 10.1186/1748-5908-4-54] [Medline: 19674440]
24. 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(85):2825-2830.
25. Hayes RJ, Moulton LH. Cluster randomised trials. Boca Raton, FL: Chapman and Hall/CRC; 2009.
26. Lin J, Gamalo-Siebers M, Tiwari R. Propensity score matched augmented controls in randomized clinical trials: a case study. Pharm Stat 2018;17(5):629-647. [doi: 10.1002/pst.1879] [Medline: 30066459]
27. Governance-document: Nivel Zorgregistraties Eerste Lijn. Nivel. URL: https://www.nivel.nl/sites/default/files/bestanden/governance-document-nzr-3.pdf [accessed 2022-03-24]
28. Cohen J. Statistical power analysis for the behavioral sciences. Cambridge, MA: Academic Press; 1969.
29. Ranganathan P, Aggarwal R, Pramesh CS. Common pitfalls in statistical analysis: odds versus risk. Perspect Clin Res 2015;6(4):222-224 [FREE Full text] [doi: 10.4103/2229-3485.167092] [Medline: 26623395]
30. Nielen MM, Spronk I, Davids R, Korevaar JC, Poos R, Hoeymans N, et al. Estimating morbidity rates based on routine electronic health records in primary care: observational study. JMIR Med Inform 2019;7(3):e11929 [FREE Full text] [doi: 10.2196/11929] [Medline: 31350839]
31. Beauchemin M, Murray MT, Sung L, Hershman DL, Weng C, Schnall R. Clinical decision support for therapeutic decision-making in cancer: a systematic review. Int J Med Inform 2019;130:103940 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.07.019] [Medline: 31450082]
32. Laka M, Milazzo A, Merlin T. Can evidence-based decision support tools transform antibiotic management? A systematic review and meta-analyses. J Antimicrob Chemother 2020;75(5):1099-1111. [doi: 10.1093/jac/dkz543] [Medline: 31960021]
33. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA Open 2020;3(2):167-172 [FREE Full text] [doi: 10.1093/jamiaopen/ooz046] [Medline: 32734155]
34. Riley P. Three pitfalls to avoid in machine learning. Nature 2019;572(7767):27-29. [doi: 10.1038/d41586-019-02307-y] [Medline: 31363197]

35. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. J Biomed Inform 2018;78:134-143 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.005] [Medline: 29246790]

Abbreviations

CDSS: clinical decision support system
EHR: electronic health record
GP: general practitioner
LUMC: Leiden University Medical Center
ML: machine learning
NHG: Dutch College of General Practitioners
UTI: urinary tract infection

Edited by C Lovis; submitted 07.02.21; peer-reviewed by M Campos, R Gunnarsson , J Walsh; comments to author 02.04.21; revised version received 28.05.21; accepted 13.02.22; published 04.05.22.

Please cite as:
Herter WE, Khuc J, Cinà G, Knottnerus BJ, Numans ME, Wiewel MA, Bonten TN, de Bruin DP, van Esch T, Chavannes NH, Verheij RA
Impact of a Machine Learning–Based Decision Support System for Urinary Tract Infections: Prospective Observational Study in 36 Primary Care Practices
JMIR Med Inform 2022;10(5):e27795
URL: https://medinform.jmir.org/2022/5/e27795
doi:10.2196/27795
PMID:35507396

Original Paper

Transformer- and Generative Adversarial Network–Based Inpatient Traditional Chinese Medicine Prescription Recommendation: Development Study

Hong Zhang[1*], MSc; Jiajun Zhang[2*], PhD; Wandong Ni[3], PhD; Youlin Jiang[1], MSc; Kunjing Liu[1], BSc; Daying Sun[4], PhD; Jing Li[1], MSc

[1]Guanganmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China

[2]School of Electronic Information Engineering, Wuxi University, Wuxi, China

[3]Physician Qualification Program, Certification Center of Traditional Chinese Medicine, State Administration of Traditional Chinese Medicine, Beijing, China

[4]School of Electronic Engineering and Optoelectronic Technology, Nanjing University of Science and Technology, Nanjing, China

[*]these authors contributed equally

Corresponding Author:
Wandong Ni, PhD
Physician Qualification Program
Certification Center of Traditional Chinese Medicine
State Administration of Traditional Chinese Medicine
No 5 Beixian Ge Road
Xicheng District
Beijing, 100053
China
Phone: 86 13311127900
Email: 2592967878@qq.com

Abstract

Background: Traditional Chinese medicine (TCM) practitioners usually follow a 4-step evaluation process during patient diagnosis: observation, auscultation, olfaction, inquiry, pulse feeling, and palpation. The information gathered in this process, along with laboratory test results and other measurements such as vital signs, is recorded in the patient's electronic health record (EHR). In fact, all the information needed to make a treatment plan is contained in the EHR; however, only a seasoned TCM physician could use this information well to make a good treatment plan as the reasoning process is very complicated, and it takes years of practice for a medical graduate to master the reasoning skill. In this digital medicine era, with a deluge of medical data, ever-increasing computing power, and more advanced artificial neural network models, it is not only desirable but also readily possible for a computerized system to mimic the decision-making process of a TCM physician.

Objective: This study aims to develop an assistive tool that can predict prescriptions for inpatients in a hospital based on patients' clinical EHRs.

Methods: Clinical health records containing medical histories, as well as current symptoms and diagnosis information, were used to train a transformer-based neural network model using the corresponding physician's prescriptions as the target. This was accomplished by extracting relevant information, such as the patient's current illness, medicines taken, nursing care given, vital signs, examinations, and laboratory results from the patient's EHRs. The obtained information was then sorted chronologically to produce a sequence of data for the patient. These time sequence data were then used as input to a modified transformer network, which was chosen as a prescription prediction model. The output of the model was the prescription for the patient. The ultimate goal is for this tool to generate a prescription that matches what an expert TCM physician would prescribe. To alleviate the issue of overfitting, a generative adversarial network was used to augment the training sample data set by generating noise-added samples from the original training samples.

Results: In total, 21,295 copies of inpatient electronic medical records from Guang'anmen Hospital were used in this study. These records were generated between January 2017 and December 2018, covering 6352 types of medicines. These medicines were sorted into 819 types of first-category medicines based on their class relationships. As shown by the test results, the

performance of a fully trained transformer model can have an average precision rate of 80.58% and an average recall rate of 68.49%.

Conclusions: As shown by the preliminary test results, the transformer-based TCM prescription recommendation model outperformed the existing conventional methods. The extra training samples generated by the generative adversarial network help to overcome the overfitting issue, leading to further improved recall and precision rates.

(JMIR Med Inform 2022;10(5):e35239) doi:10.2196/35239

KEYWORDS

traditional Chinese medicine; transformer; generative adversary networks; electronic health records; artificial intelligence; natural language processing; machine learning; word2Vec

Introduction

The widespread use of electronic health record (EHR) systems has led to the explosive growth of digitized health care data. As the amount and complexity of data grow, medical analysis and decision-making become increasingly time-consuming and error prone. In reality, a human physician cannot fully use all the available information at his or her disposal in a timely fashion. Therefore, harnessing the information contained in EHR data, most of which is in textual form, is critical for driving innovation research, improving health care quality, and reducing costs. Natural language processing (NLP) is essential for transforming relevant information sequestered in freestyle texts into structured data for further computerized processing. The development of a predictive model with EHR data was motivated by the desire to offer a medication-oriented decision support tool to clinical health care providers. To build such a predictive model, we used NLP techniques to convert a patient's EHR data into a representation, which then becomes the input to a deep learning model to predict medical events, such as medication orders.

Biomedical NLP has experienced great progress in the past 30 years [1,2] and has become especially active in recent years [3]. Previously, EHR data were analyzed using traditional machine learning and statistical techniques such as logistic regression, support vector machine, and random forest [4]. However, in recent years, as reviewed in the studies by Shickel et al [5], Sheikhalishahi et al [6], and Miotto et al [7], many research efforts have been devoted to the application of deep learning techniques to EHR data for clinical informatics tasks. Autoencoders have been used by researchers [8] to predict a specific set of diagnoses. A long short-term memory (LSTM) sequence model [9] was trained to provide patient-specific and time-specific predictions of medication orders for patients who are hospitalized [10]. A convolutional neural network (CNN) model was used to predict discharge medications using the information available at admission [11]. Numerous articles were surveyed in the study by Goldstein et al [12] regarding the development of a risk prediction model using EHR data. A comprehensive study on applying deep learning techniques to EHR data for a variety of prediction problems was reported in the study by Rajkomar et al [13]. Recurrent neural networks were successfully trained using EHR data to detect medical events [14-16].

The research on applying artificial intelligence in traditional Chinese medicine (TCM) has been very active in the past decade [17,18]. Data mining techniques have been used for TCM syndrome modeling and prescription recommendation for diabetes [19]. The PageRank algorithm [20] was modified and applied to TCM prescription recommendations [21]. In our previous work [17], a CNN was used to predict TCM diseases, and XGBoost, along with other neural networks, was used to predict TCM syndromes. Following the sequence-to-sequence paradigm, researchers from Peking University used bidirectional gated recurrent neural networks to generate TCM prescriptions from symptom descriptions [22]. They proposed a coverage mechanism along with a soft loss function as a remedy for the repetition problem they encountered. However, the requirement of curated descriptions of symptoms as inputs hinders the practicality of this approach. Ideally, the model generates TCM prescriptions directly from raw EHR data, similar to how a human TCM physician conducts deductive reasoning.

Generating prescriptions from raw EHR data typically comprises 2 parts. The first part uses biomedical NLP [3] techniques to extract relevant information used by a human physician to form a feature representation [23]. The second part uses deep learning techniques [7] to map this feature representation into a prescription order.

The primary task of biomedical NLP is to extract relevant information from clinical narratives written in free-form text and store the gathered information as structured data. Numerous deep learning techniques [24-26], such as bidirectional LSTM (BiLSTM), have been used in the biomedical NLP field. Both BiLSTM conditional random field (CRF) and transformer CRF have been used for named entity recognition (NER) of EHR notes written in Chinese [27,28]. The recognized entities are then formed into distinct tokens. Then, the feature representation of a patient's EHR data becomes a sequence of tokens. The tokens are then converted into real-valued multidimensional vectors using word embedding techniques [29].

The purpose of this study was to develop an assistive tool that can prescribe TCM prescriptions for inpatients in a hospital based on the patient's clinical EHRs. The predictive model for TCM prescription generation is based on a sequence-transducing model called the transformer [30]. This model is entirely based on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multihead self-attention. The training used in this predictive model was supervised training with human-authored prescriptions contained in the EHR data set as the training targets. Furthermore, a generative adversarial network (GAN) [31] model was designed to augment

the training set to further enhance the overall system performance by reducing the effects of overfitting.

Methods

This section is arranged as follows: the overall system architecture is briefly described; then, each constituent subsystem, which may comprise some functional blocks, is introduced; finally, the training process is described in the *Training* subsection, where a GAN model was used to generate noise-added samples from the original samples.

System Overview

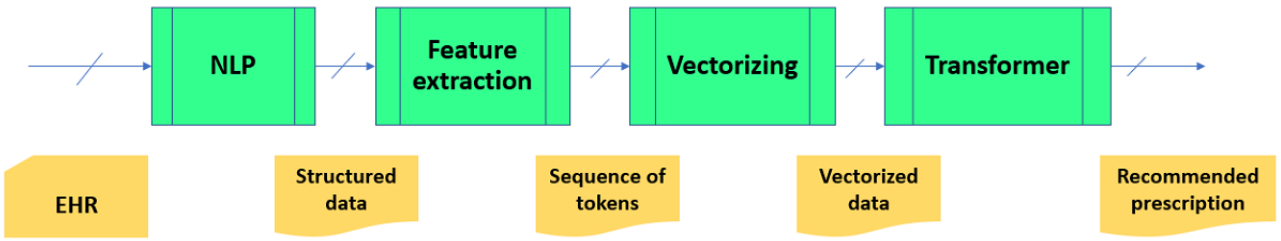
Hospitals and medical institutes in China are rapidly moving toward standardizing their EHRs to conform to the regulations and specifications issued by the Ministry of Health of the People's Republic of China [32-34]. A standard EHR document for a patient may contain up to 53 parts, depending on the patient's situation. These may include the following:

- A first page record containing the patient's basic personal information, such as sex, age, occupation, and marital status
- An admission record containing the description of a patient's illness upon admission to the hospital, including chief complaints, medical history, and family medical history
- A laboratory tests record containing the list of tests and the corresponding results
- A nursing record containing nurse notes of the patient's condition, treatments taken and nursing care taken, body temperatures and vital signs taken, and physician's orders
- A treatment procedure record containing the entire in-hospital diagnosis and treatment process and any changes to the patient's illness or illnesses

A high-level block diagram of the proposed system is shown in Figure 1. The system comprises 4 subsystems: the NLP subsystem, the feature extraction subsystem, the vectorization subsystem, and the prescription prediction subsystem. The NLP subsystem processes the EHR file and produces structured data, which in turn are processed by the feature extraction subsystem to extract relevant clinical information for prescription prediction. The vectorization subsystem maps the sequence of tokens written in Chinese characters to digital numbers, presented as a vector in a multidimensional space. The prescription prediction subsystem, which is a transformer-based deep learning model, automatically generates a prescription based on input vector data. Together, the first 3 subsystems accomplish the task of extracting relevant information from an EHR file to form input variables for the prediction model. Similar representation learning operations were described in our previous paper [17].

In short, NLP normalizes the raw EHR data, the feature extractor converts the normalized data into a sequence of tokens, the vectorization subsystem maps the tokens into vectors of real numbers, and the predictive model performs the reasoning process to produce a prescription.

Figure 1. Block diagram of the prescription generation system. EHR: electronic health record; NLP: natural language processing.



The NLP Subsystem

This subsystem is responsible for generating structured data from original EHR documents. The internal block diagram of the subsystem is shown in Figure 2. There are 3 functional blocks in this subsystem: the preprocessing block, NER block, and British Medical Journal block.

The preprocessing block cleans the raw EHR document by removing pictures and unusable components. This ensures the completeness and accuracy of the electronic medical records. Electronic medical records with incomplete or inconsistent information are discarded.

After the initial cleaning, the content of the EHR file is then divided into distinct sections. For example, the admission record is divided into sections of chief complaints, medical history, and others. Then, all the resultant sections are sorted, formatted, and subsequently fed to the NER block.

Only a small part of the EHR document is in a fixed format, and the remainder is in unstructured freestyle narratives. For fixed-format texts, a script is used to extract named entities to form structured data.

For freestyle narratives, a functional block called entity recognition is used to extract named entities to form structured data entries. The NER block is implemented using a BiLSTM network with CRF (BiLSTM-CRF) [24].

Then, the extracted named entities such as symptoms, illness, medicine, examinations, and tests are further standardized according to a Chinese version of the British Medical Journal Best Practice knowledge base.

Figure 3 shows an example of the processing result, where the admission record of a raw EHR note is converted into structured data, with the marked words being named entities.

Figure 2. Block diagram of the named entity recognition subsystem. BiLSTM: bidirectional long short-term memory; EMR: electronic medical record; BiLSTM-CRF: Bidirectional long short term memory – conditional random fields; BMJ: British Medical Journal.

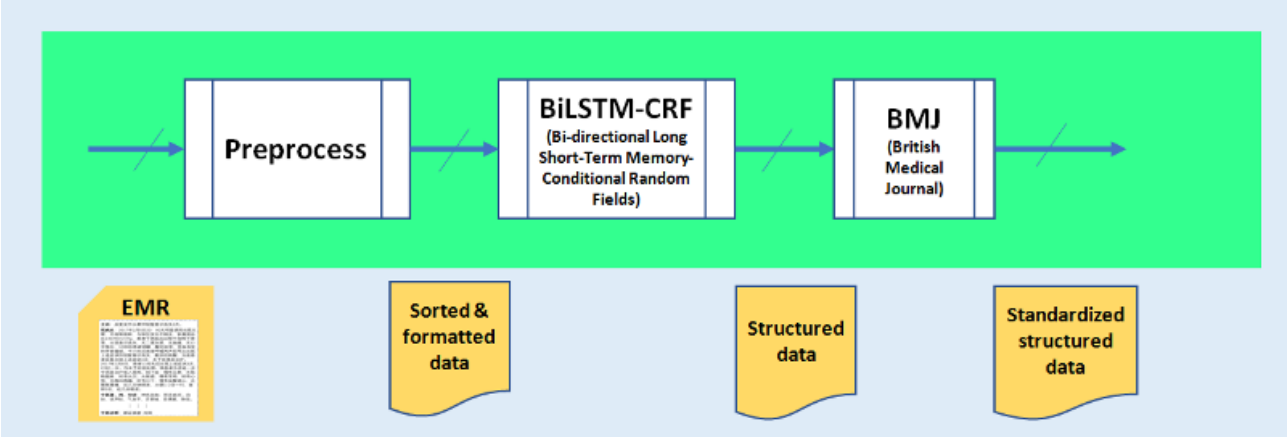
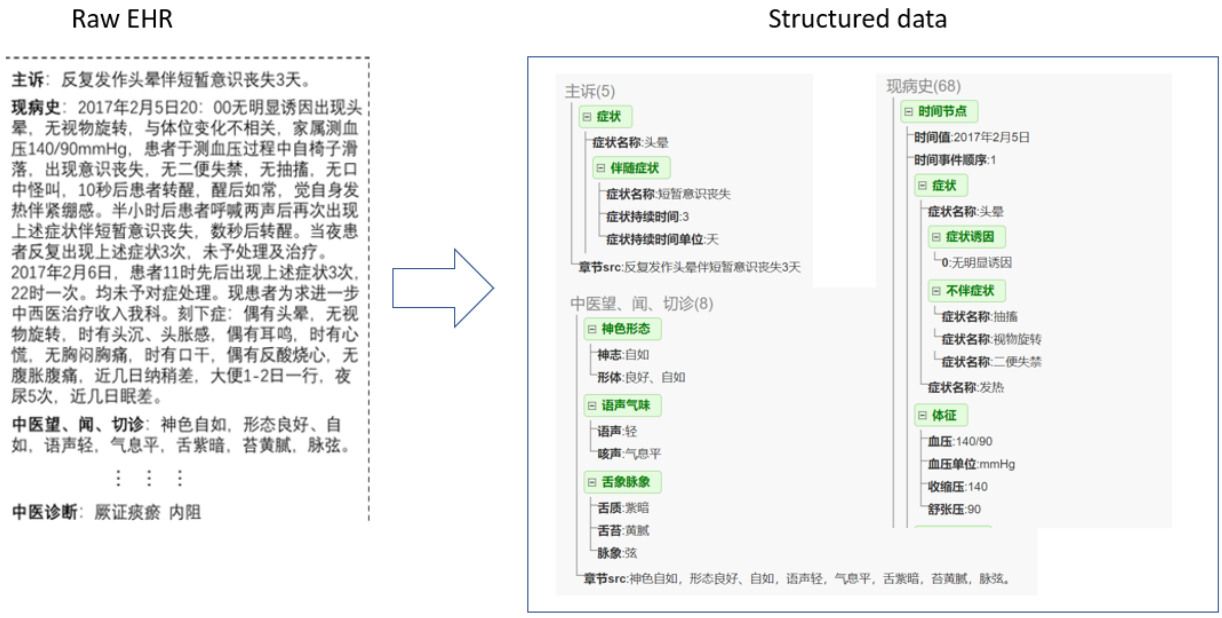


Figure 3. Example of converting a freestyle narrative into structured data. EHR: electronic health record.



The Feature Extraction Subsystem

To effectively mimic the reasoning process conducted by a human physician, accurate and relevant input variables must be chosen properly. These variables should represent the complete set of factors that a human physician should take into consideration when making treatment decisions. Textbox 1 summarizes the predominant factors that TCM experts consider when making treatment decisions.

The feature extraction subsystem extracts the aforementioned key features from the standardized structured data to form a sequence of tokens. Figure 4 shows an example of this feature extraction, in which a sequence of tokens is generated from structured data.

Textbox 1. Text type and the content to extract.

Demography

- Sex, age, height, weight, and BMI

Chief complaints

- Symptoms and signs

Recent medical history

- Symptoms, signs, and general information

Past medical history

- Past illness and medicines taken

Present illness

- Tongue coating and pulses

Body check

- Vital signs

Treatment process records

- Current illness situation and treatment plan

Physician's orders

- Prescriptions

Nursing notes

- Vital signs and medication records

Examination reports

- Examination items and findings

Laboratory reports

- Items tested and qualitative and quantitative test results

Figure 4. Example of converting structured data into a sequence of tokens.

['头晕', '短暂性意识丧失', '3天', '无明显诱因', '头晕', '无视物旋转', '体位变化', '不相关', '血压检查', '140/90mmhg', '血压检查', '意识丧失', '无大小便失禁', '无抽搐', '10秒后', '转醒', '发热', '紧绷感', '半小时后', '短暂性意识丧失', '数秒后', '转醒', '偶有', '头晕', '无视物旋转', '头部沉重感', '头部发胀', '偶有', '耳鸣', '时有', '心慌', '无胸闷', '无胸部疼痛', '时有', '口干', '偶有', '反酸', '烧心', '无腹胀', '无腹部疼痛', '近几日', '眠差', '神色自如', '形态', '良好', '自如', '语声轻', '气息平', '舌紫暗', '苔黄腻', '脉弦']

The Vectorization Subsystem

Overview

Until this point, all medical information needed to make a treatment decision was encapsulated in textual data expressed in Chinese characters. To be used by the deep learning network—the Transformer—the information must be mapped into a digital variable. In this vectorization process, a Chinese word or phrase is represented as a real-valued vector in multidimensional feature space. This section explains how tokenized features are further processed through word embedding.

Training the Word Embedding Model

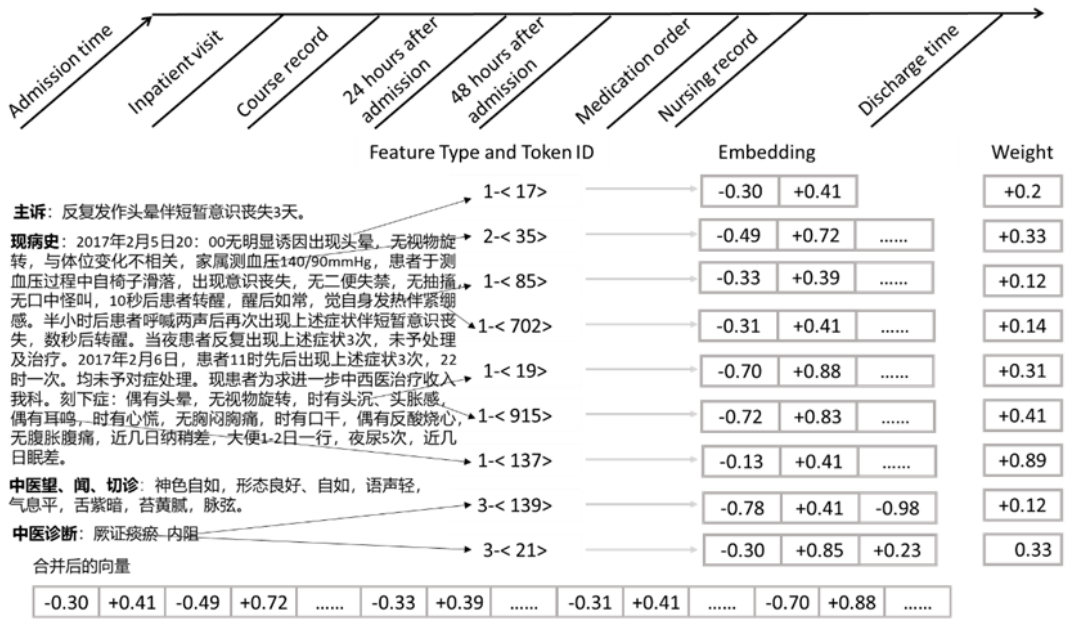
The corpus was a collection of 102,596 electronic medical records from Guang'anmen Hospital and other hospitals. The *Jieba* tokenizer was used to perform tokenization. The open-source modeling tool *Gensim* was used to train the word2vec [29] model with the following major parameters: *min_count*=2, *vector_size*=100, *window*=5, *sg*=1, *hs*=1, and *epochs*=50. The Skip-Gram model was used, as indicated by the parameters. Each word was represented by a real-valued vector of 100 dimensions.

Vectorization

Once the word embedding model is trained, each token is represented by a 100-dimension vector. For each word in the input sequence, a unique identifier is assigned using a numerical-type value expressed as a name-value-unit before another unique identifier is assigned. Once all tokens are converted into vectors, the vectors are then concatenated to form a single vector variable, which then serves as the input to the transformer.

The NLP, feature extraction, and vectorization subsystems together accomplish the task of feature learning by converting an EHR document into a multidimensional real-valued vector. Figure 5 shows an example of mapping from EHR text to word vectors.

Figure 5. Illustration of converting electronic health record text to word vectors.



The Transformer Subsystem

The transformer subsystem is responsible for recommending a prescription for every given input embedding, as shown in Figure 6. The subsystem is described in the following paragraphs.

Input embedding is a vector of *max_num_tokens*× *vector_size* dimensions. For example, *max_num_tokens*=759 and *vector_size*=100. Zero padding is used if the number of tokens in a sequence is smaller than *max_num_tokens*. Conversely, if the number of tokens in a sequence is larger than *max_num_tokens*, the number of tokens is capped at *max_num_tokens* by dropping off tokens corresponding to the oldest time stamp with respect to the current prescription generation time. The input embedding sample is first added to the position vector of the same size, becoming the input to the first encoder.

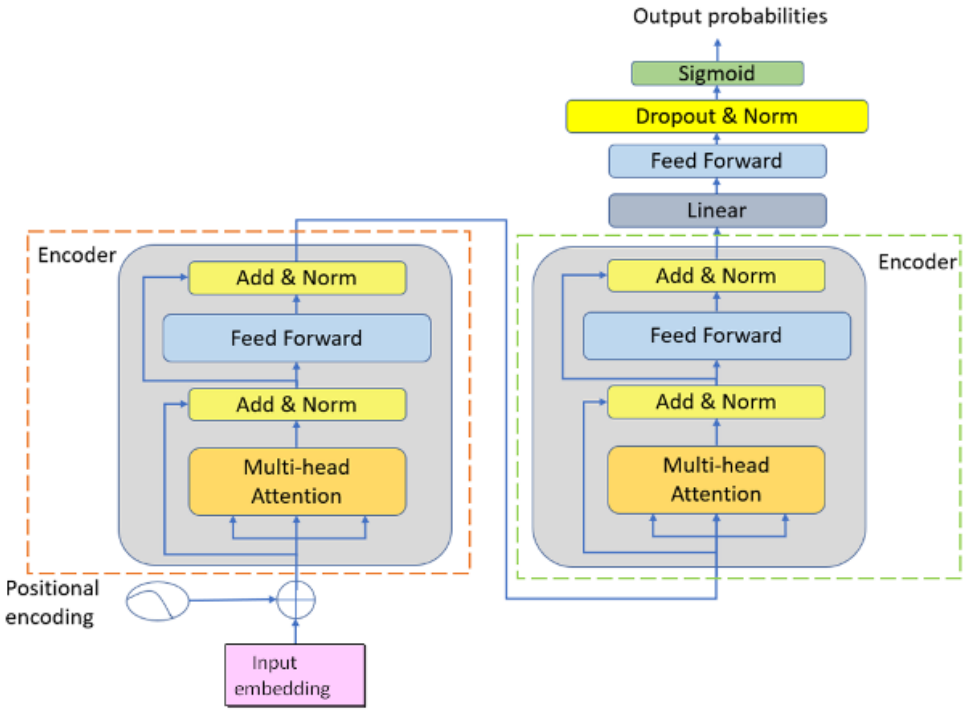
The main body of the subsystem comprises 2 identical cascaded transformer encoders. Unlike the encoder of the original transformer [30], which comprises 6 identical layers, the encoder used in this research had only 1 layer with 4 sublayers. The first was a multihead self-attention layer with *Multi_heads*=4 and *head_dim*=8. The second was a residual layer of 100 neurons with normalization. The third was a simple, position-wise, fully connected feedforward network of 2048 neurons. The fourth was a residual layer of 100 neurons with normalization.

The second encoder was followed by a linear layer, a feedforward layer of 2048 neurons, a hidden layer, and an output layer, as shown in Figure 6. The output layer comprised 819 neurons with a sigmoid activation function. Each of the 819

neurons corresponded to an herbal ingredient. The hidden layer comprised 128 neurons with a dropout mechanism and normalization. The dropout rate was set to 0.4740. The purpose of this hidden layer was to prevent overfitting.

The final result from the output layer was a list of probabilities for the 819 drug ingredients, valued between 0 and 1. The recommended prescription was then obtained by setting a threshold for these probabilities.

Figure 6. The transformer subsystem.



Training

Training the Transformer

Training of the transformer is a supervised learning process. The input is a real-valued vector representation of a patient's EHR, and the output is the prescription. The learning goal is for a machine-generated prescription to match the medical order prescribed by a human physician.

Augmenting the Training Data

To alleviate the overfitting effect of the proposed prediction model, a GAN [31] network was used to augment the training data set. Following the fundamental idea of the GAN network, the generative model G is trained to represent the distribution of the original training data set, and the discriminative model D is trained to detect whether the sample originates from the original sample set or from the output of the generative model.

During the training phase, the entire system looks like that shown in Figure 7. For every original training sample, there is a noise-added sample. The use of a GAN in this system effectively doubled the number of training samples.

The internal structure of our GAN network was designed as shown in Figure 8. Generator G comprises 2 identical LSTM layers, each with a size of 279. Each LSTM layer is followed by a normalization layer with a residual connection. The input to the discriminator G could be either an original word embedding sample or a noise-added sample generated by the generator G. The discriminator D comprises an LSTM layer with a size of 279, a residual and normalization layer with a size of 100, and a full connection layer with a size of 256. Finally, the discriminator D outputs a binary value using a sigmoid function.

We followed a typical GAN network training procedure [31] to train the GAN subsystem, simultaneously training the discriminator and generator. The discriminator and generator alternate in their training until a Nash equilibrium is reached.

The generator first produces a *batch_size* noise-added EHR, embedding samples with randomly initialized coefficients of the generator network. These samples are concatenated with the original noise-free EHR embedding samples to form (2×*batch_size*) embedding samples, each with *max_num_tokens*×*vector_size* real values. For example, we can have *batch_size*=500, *max_num_tokens*=560, *vector_size*=100. These (2×*batch_size*) samples were used as inputs to the discriminator. For every input sample, an output label indicates whether the sample is from the true original embedding or from the generator. The discriminator network was trained using a backpropagation algorithm with the objective of minimizing the prediction error. The training of the discriminator is halted when the binary cross-entropy loss function stops decreasing. The discriminator training is then temporarily halted to yield to the generator training.

To train the generator, all network coefficients of the discriminator must be frozen. The discriminator now works in tandem with the generator during generator training. The generator produces *batch_size* noise-added embedding samples,

and for every sample, the discriminator outputs a prediction. The generator updates its parameters using a backpropagation algorithm based on the discriminator output. The training of the generator is halted when the binary cross-entropy loss function stops increasing. The generator training is then temporarily halted to yield the discriminator training.

The aforementioned discriminator and generator training processes together form 1 training epoch. The entire GAN network training is accomplished through several epochs. The training stops when a Nash equilibrium is reached.

The entire training process is illustrated using the Python pseudocode included in Multimedia Appendix 1.

Figure 7. Block diagram of the predictive modeling system during the training phase. EHR: electronic health record; GAN: generative adversarial network; NLP: natural language processing.

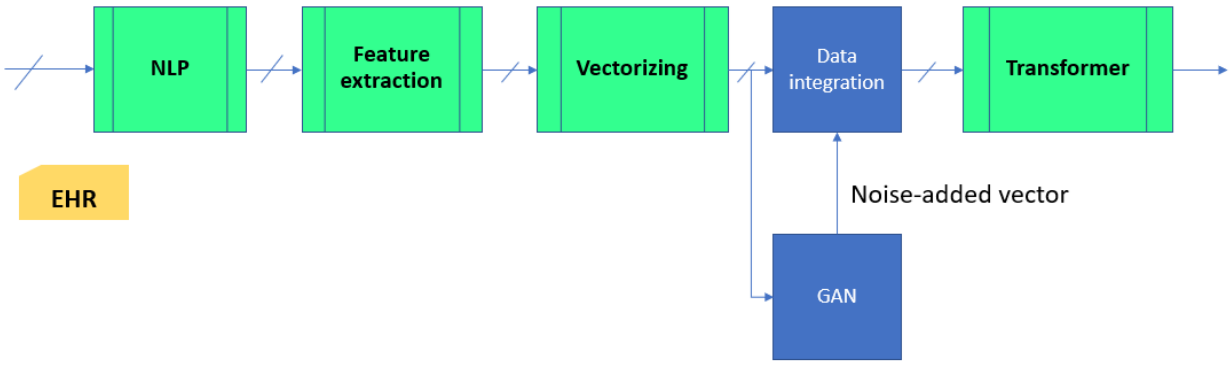
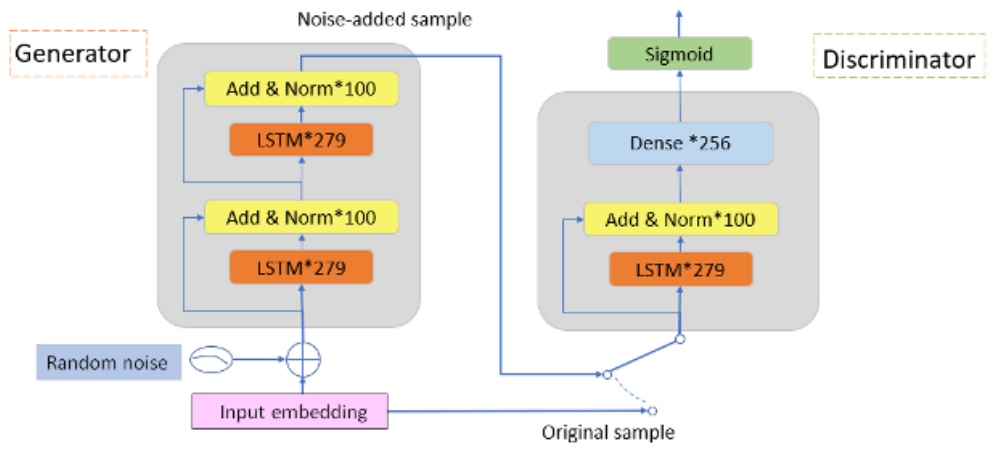


Figure 8. The internal structure of the generative adversarial network subsystem. LSTM: long short-term memory; *size of the neural network used in that layer.



Ethics Approval

This study received institutional review board review through Guanganmen Hospital Ethic Committee (SQ2017YFGX 060073).

Results

Data Set

EHRs generated in Guang'anmen Hospital between January 1, 2017, and December 31, 2018, were used as the data set in this study. Initially, there were 27,846 copies of EHR notes, out of which 6551 (23.53%) copies were discarded because of quality control. An EHR note should be discarded if it satisfies one of the following conditions:

- The note is incomplete for missing certain basic pages.
- The note contains inconsistent information.
- The note does not use standard descriptions.
- The note contains special EHR circumstances such as chemotherapy, after an operation, and removal of fracture settings.

Evaluation Metrics

The data set contained 6352 drug varieties. A complete TCM prescription includes drug ingredients, dosages, and decoction preparation instructions. It is still very challenging, if not impossible, for a machine to generate such a complete TCM prescription. At our current stage of research, we focus only on the drug ingredients of a prescription.

Judging whether the 2 TCM prescriptions are the same is often not straightforward, given the distinctive nature of TCM [35]. Often, 2 different herbs may have the same medical effect. When a TCM physician prescribes a medication order, he or she often has multiple choices at hand for herbal ingredients. As a result, the 2 TCM physicians may prescribe different herbs for the same patient with the same diagnosed condition. Therefore, it is necessary to have a unified method of evaluating machine-generated prescriptions. To this end, we need a higher

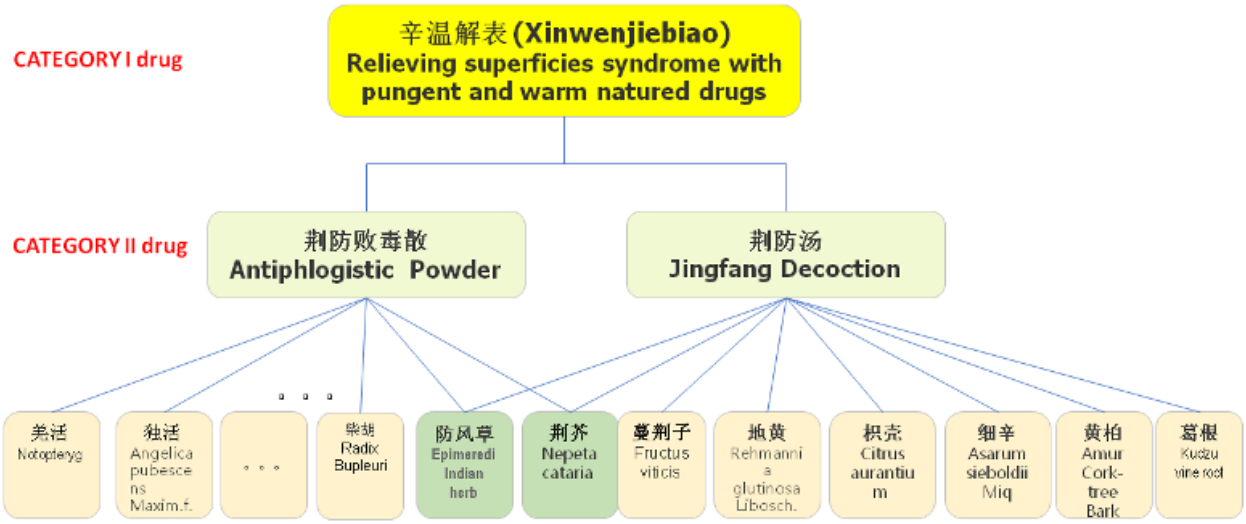
level of abstraction. Figure 9 shows an example of the organization of TCM drugs. In this example, 2 TCM drugs (antiphlogistic powder and Jingfang decoction) have different herbal ingredients but belong to the same parent drug category and have the same medical treatment effect. In our research, we concluded that the recommended drug should be considered a correct recommendation as long as the recommended drug belongs to the same parent category as that of the human-authored prescription.

To quantitatively evaluate the performance of the transformer-based deep learning model, we compared the prescription generated by the machine with that prescribed by a human physician. Here, we used the metrics of *precision rate* and *recall rate*, which we based on 3 variables. True positive (TP) is defined as the number of drugs that exist in the physician's prescription and also exist in the machine's prescription. False positive (FP) is the number of drugs that do not exist in the physician's prescription but exist in the machine's prescription. False negative (FN) is defined as the number of drugs that exist in the physician's prescription but not in the machine's prescription. With these definitions, we defined the precision and recall rates as follows:

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

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

Figure 9. Classification of herbal drugs.



Hyperparameter Tuning With GridSearchCV

The data set was divided into training and test sets, with the training set comprising 90% of the data set and the test set comprising the rest. The model was trained using a 10-fold cross-validation method; that is, the training set was randomly split into 10 folds, with the model being trained 10 times. During each of the 10 training times, the hyperparameters were tuned using the GridSearchCV method. Each training resulted in a set of hyperparameters, with the ultimate hyperparameters being the average of these 10 sets of parameters.

The values of the hyperparameters of the transformer network model have a great influence on the accuracy of the model. The optimal values of these parameters were determined through iterations using the grid search method. The sparse characters of each type were embedded into a d-dimensional embedding layer. Then, all vectors were combined using a new method: vectors of the same type and time were averaged using the weights of self-learning.

The model was optimized using a minimal log loss. Many regularization methods were used, such as the vector loss rate and the embedded layer loss rate. In addition, small-scale L2 weight punishment was used, which increased the punishment for large weights. The training batch size was chosen as 128, placing sentences with similar sizes into the same batch. Each batch contained approximately 12,000 words. Finally, the multilabel task was processed using an Adam function. For multilabel tasks, the input with the last time stamp was multiplied with the special end of sequence embedding. The training was executed using the Kears framework on a server with 8 NVIDIA P100 graphics processing unit. The fine-tuned hyperparameters along with their respective ranges are shown in Table 1.

Table 1. Some hyperparameters of the model.

Hyperparameters	Values	Parameter range
Gradient	0.1245	(0.1, 0.5, 1.0, 1.1)
Attention heads	4	(4, 8)
Vector loss rate	0.4410	(0.25, 0.35, 0.5)
Hidden layer loss rate	0.4740	(0.25, 0.35, 0.5)
Learning rate	0.4375	(0, 1)
L2 punishment rate	0.000001566	(0, 0.01)

Experimental Results

To intuitively explain our experimental results, we start with a concrete example that illustrates how EHR notes lead to prescription orders. An example of this is shown in Figure 10. The left side shows a snapshot of the patient's EHR. On the right side is a table showing a side-by-side comparison between a human-authored order and the prescription generated by our model. The physician's order contains 12 ingredients, whereas the model's order has 11. The first 5 ingredients are identical on both sides. The sixth ingredient from each side is the same, although they have different Chinese names. This is because the physician used a nickname for the herb. The remaining ingredients differ not only in name but also in substance. However, these 2 orders are still considered equivalent so far as the medical treatment effect is concerned. This is because in TCM terminology, a diagnosis must conclude with the name of the disease (illness) and a list of syndromes [17]. In this particular case, the diagnosed disease is *emaciation-thirst*, with the primary syndrome being *kidney and liver deficiency* and the secondary syndrome being *dampness and stasis*. The first 6 herbal ingredients target the primary syndrome. The remaining ingredients in each prescription are for the treatment of the secondary syndrome called *dampness and stasis*. As these 2 orders are only slightly different in their ingredients for treating secondary syndrome, they are treated as the same prescription in our research.

To further explain this prescription comparison, we present another picture, as shown in Figure 11. The physician's order is called *Qiju Dihuang pill*, and the model's order is called *Liuwei Dihuang pill*. They are category II prescriptions that belong to the same parent category TCM prescription called *nourishing liver and kidney*. They differ only in how to dispel dampness and resolve phlegm to address only the secondary syndrome.

To evaluate the performance of the transformer-based predictive model, we first conducted model training using only the original samples, purposefully excluding the noise-added samples. The results are described in the following paragraphs.

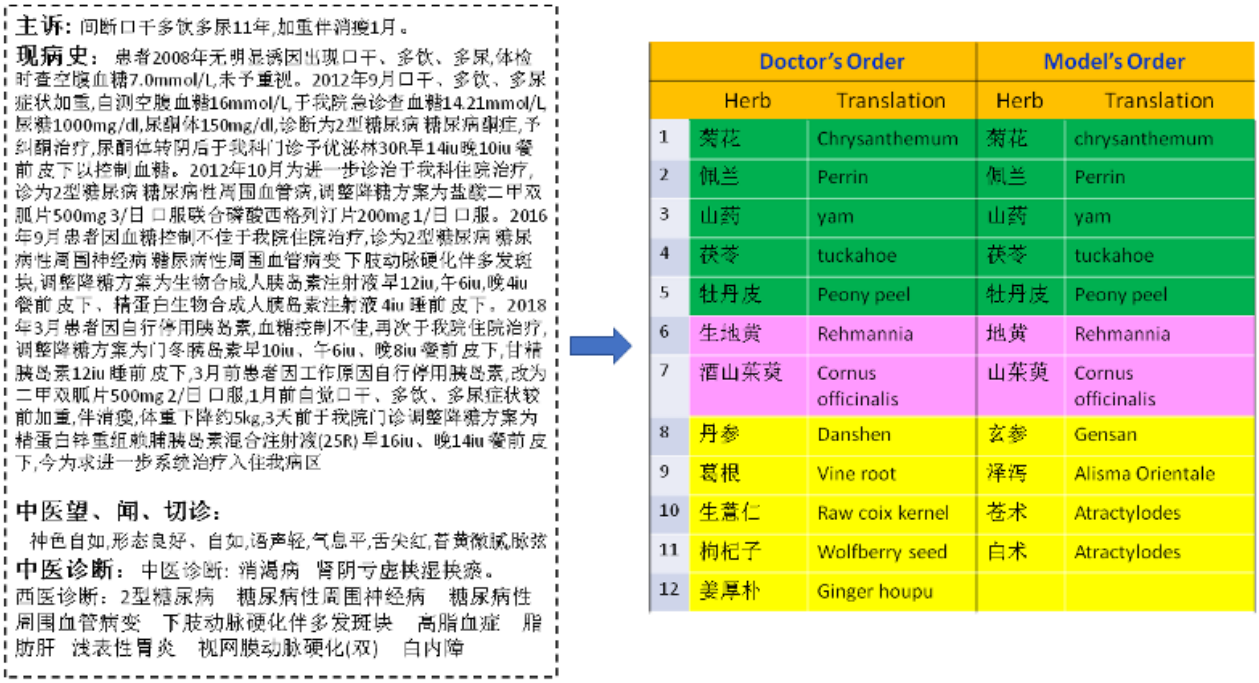
On the basis of the time sequences, the system produced prescription recommendations at admission, 24 hours after admission, 48 hours after admission, 3 days after admission, and 1 week after admission. The test results are shown in Table 2.

From Table 2, we first observe that the precision and recall rates obtained from the training data set are higher than their respective counterparts from the test data set. This is understandable as the model has seen the samples from the training data set before but not from the test data set. The second observation is that as time progresses, both the precision and recall rates improve. After admission, at each subsequent medication order time, more relevant information is collected, and the prediction becomes more accurate. Although the number of feature tokens was <260 for 98% of the patients at the time of admission, this number increased to 296 in 24 hours, 333 in 48 hours, 366 in 72 hours, and 759 in 7 days. In our experiment, we set *max_num_tokens*=759. This means that when the number of feature tokens was <759, zero padding was used, and clipping was used when there were >759 feature tokens. Selecting the proper value for *max_num_tokens* is important for balancing the trade-off between overall system performance and computational efficiency. If the value is too large, training and inferencing will consume too much computation horsepower. If the value is too small, then some critical information gathered at admission will be lost because of clipping, leading to reduced precision and recall rates for prescription predictions at a time that is far from the admission time (eg, 2 weeks after admission).

The second set of experimental results was obtained using more training samples to train the predictive model. The size of the training data set was doubled, as for every training sample, a noise-added sample was generated by the GAN network. The precision and recall rates are listed in Table 3.

As can be seen in Table 3, both the precision and recall rates consistently improved by a noticeable margin. The results convincingly prove that inserting noise-added training samples generated by the GAN module can effectively overcome the overfitting issue, leading to better prediction performance.

Figure 10. Side-by-side comparison of physician's order versus model's order.



主诉：间断口干多饮多尿11年,加重伴消瘦1月。

现病史：患者2008年无明显诱因出现口干、多饮、多尿,体检时查空腹血糖7.0mmol/L,未予重视。2012年9月口干、多饮、多尿症状加重,自测空腹血糖16mmol/L,于我院急诊查血糖14.21mmol/L,尿糖1000mg/dl,尿酮体150mg/dl,诊断为2型糖尿病 糖尿病酮症,予纠酮治疗,尿酮体转阴后于我科门诊予优泌林30R早14iu晚10iu 餐前皮下以控制血糖。2012年10月为进一步诊治于我科住院治疗,诊为2型糖尿病 糖尿病性周围血管病,调整降糖方案为盐酸二甲双胍片500mg 3/日 口服联合磷酸西格列汀片200mg 1/日 口服。2016年9月患者因血糖控制不佳于我院住院治疗,诊为2型糖尿病 糖尿病性周围神经病 糖尿病性周围血管病变 下肢动脉硬化伴多发斑块,调整降糖方案为生物合成人胰岛素注射液早12iu,午6iu,晚4iu 餐前皮下、精蛋白生物合成人胰岛素注射液 4iu 睡前皮下。2018年3月患者因自行停用胰岛素,血糖控制不佳,再次于我院住院治疗,调整降糖方案为门冬胰岛素早10iu、午6iu、晚8iu 餐前皮下,甘精胰岛素12iu 睡前皮下,3月前患者因工作原因自行停用胰岛素,改为二甲双胍片500mg 2/日 口服,1月前自觉口干、多饮、多尿症状较前加重,伴消瘦,体重下降约5kg,3天前于我院门诊调整降糖方案为精蛋白锌重组赖脯胰岛素混合注射液(25R)早16iu、晚14iu 餐前皮下,今为求进一步系统治疗入住我病区

中医望、闻、切诊：

神色自如,形态良好、自如,语声轻,气息平,舌尖红,苔黄微腻,脉弦

中医诊断：中医诊断：消渴病 肾阴亏虚挟湿挟瘀。

西医诊断：2型糖尿病 糖尿病性周围神经病 糖尿病性周围血管病变 下肢动脉硬化伴多发斑块 高脂血症 脂肪肝 浅表性胃炎 视网膜动脉硬化(双) 白内障

	Doctor's Order		Model's Order	
	Herb	Translation	Herb	Translation
1	菊花	Chrysanthemum	菊花	chrysanthemum
2	佩兰	Perrin	佩兰	Perrin
3	山药	yam	山药	yam
4	茯苓	tuckahoe	茯苓	tuckahoe
5	牡丹皮	Peony peel	牡丹皮	Peony peel
6	生地黄	Rehmannia	地黄	Rehmannia
7	酒山茱萸	Cornus officinalis	山茱萸	Cornus officinalis
8	丹参	Danshen	玄参	Gensan
9	葛根	Vine root	泽泻	Alisma Orientale
10	生薏仁	Raw coix kernel	苍术	Atractylodes
11	枸杞子	Wolfberry seed	白术	Atractylodes
12	姜厚朴	Ginger houpu		

Figure 11. Prescription comparison: physician's order versus model's order.

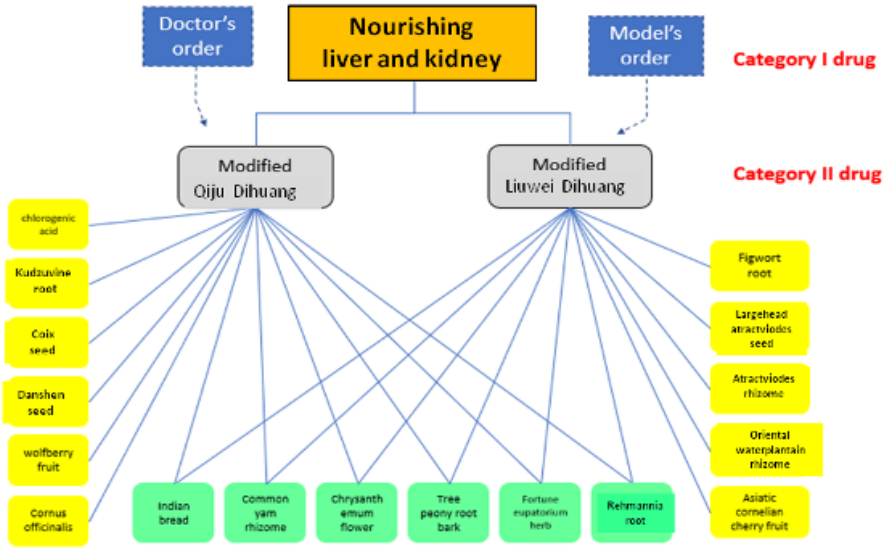


Table 2. The precision rates and recall rates with transformer only.

Time	Training set		Test set	
	Precision rate (%)	Recall rate (%)	Precision rate (%)	Precision rate (%)
Admission	81.58	69.49	73.82	61.25
In 24 hours	83.37	71.88	74.56	62.69
In 48 hours	83.92	71.26	74.81	63.04
In 3 days	85.16	73.89	76.24	65.38
In 1 week	87.02	75.17	77.94	67.15

Table 3. The precision rates and recall rates with transformer+generative adversarial network.

Time	Training set		Test set	
	Precision rate (%)	Recall rate (%)	Precision rate (%)	Recall rate (%)
Admission	82.22	70.65	80.58	68.49
In 24 hours	84.15	72.18	82.37	70.8
In 48 hours	84.32	72.56	82.92	70.26
In 3 days	87.04	75.10	85.04	74.38
In 1 week	88.91	76.79	86.82	76.23

Comparison Study

To compare the performance of our proposed model with that of existing prescription generation models, we implemented 3 other models. The CNN-based model [11] comprises a word embedding layer, a convolution layer that contains 3 filters of different sizes, a pooling layer, and a full connection layer. The output layer contains 819 neurons, equal to the number of prescribed herb varieties. The seq2seq [36] model comprises a CNN encoder and an LSTM decoder. The MedAR [37] model comprises a word embedding layer, followed by an attention layer, and finally, a RethinkNet layer to complete the multilabel classification. The learning rate was 0.001, the dropout rate was 0.8, and the optimization function was Adam. The final output layer used the sigmoid function, where all other layers used the non-linear activation function ReLU, which outputs an input x as zero if x is negative, and outputs x itself if x is larger than or equal to zero. Table 4 shows the respective precision and recall rates at admission for all 4 models in discussion. The results suggest that the proposed model has superior performance in terms of precision and recall rates.

Table 4. Performance comparison for different models.

Model	Precision rate (%)	Recall rate (%)
Convolutional neural network	47.54	31.00
Seq2seq[a]	64.02	48.74
MedAR[b]	71.46	53.08
Transformer+generative adversarial network	80.58	68.49

[a]Seq2seq: sequence to sequence model.

[b]MedAR: Medical data attention Rethink Net.

Discussion

Principal Findings

The following tasks have been finished in this research:

1. Deep learning NLP techniques were used to convert raw Chinese EHR texts into feature representations.
2. The major contribution of this study is the proposal of a transformer-based predictive modeling scheme for medication order generation from a feature representation of EHR data.
3. The secondary contribution of this study is the use of GAN to augment the training data set, leading to a noticeable performance improvement of the predictive model. Using the GAN, noise-added samples were generated to double the number of original training samples. This helped alleviate the overfitting problem, making the model more robust in terms of generalization.

Limitations

Despite the efforts made in many aspects of the diagnosis and treatment scheme recommendations, there is still much room for improvement. The training data set is still relatively small, and there may be some frequently used medicines that are not included in the training data set. The TCM prescription knowledge base is still incomplete. Some medicines do not have standard names, and no corresponding parent medicine name exists in the database. Therefore, the recommended medicine names are still the original hospital medicine names. For a multilabel prediction task, an increased number of labels will increase the difficulty of the model prediction and lower the prediction accuracy. Therefore, as a more complete knowledge base is developed, the label set will be further optimized, leading to a greater prediction accuracy of the model.

Future Work

This paper reports the preliminary research results of automated medication order generation from EHR texts for TCM inpatients who are hospitalized. The recommended medicines include Western and Chinese medicines. For Chinese medicines, only the medicine names are recommended. In the future, the dosage of the herbal ingredients, as well as the medicine preparation instructions, will be included in the recommendations. Improving the model prediction accuracy to the level of category II is also a direction for future work. Future work could expand the training data set to optimize the model.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Python pseudocode for the training of generative adversarial network with Keras Framework.
[DOCX File , 14 KB - medinform_v10i5e35239_app1.docx]

References

1. Friedman C, Rindflesch TC, Corn M. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. J Biomed Inform 2013 Oct;46(5):765-773 [FREE Full text] [doi: 10.1016/j.jbi.2013.06.004] [Medline: 23810857]
2. Hasan SA, Farri O. Clinical natural language processing with deep learning. In: Consoli S, Reforgiato Recupero D, Petković M, editors. Data Science for Healthcare: Methodologies and Applications. Cham, Switzerland: Springer; 2019:147-171.
3. Houssein EH, Mohamed RE, Ali AA. Machine learning techniques for biomedical natural language processing: a comprehensive review. IEEE Access 2021 Oct 13;9:140628-140653. [doi: 10.1109/access.2021.3119621]
4. Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press; 2012.
5. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
6. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
7. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
8. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 May 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
9. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
10. Rough K, Dai AM, Zhang K, Xue Y, Vardoulakis LM, Cui C, et al. Predicting inpatient medication orders from electronic health record data. Clin Pharmacol Ther 2020 Jul;108(1):145-154 [FREE Full text] [doi: 10.1002/cpt.1826] [Medline: 32141068]
11. Yang Y, Xie P, Gao X, Cheng C, Li C, Zhang H, et al. Predicting discharge medications at admission time based on deep learning. arXiv (forthcoming) 2017 [FREE Full text]
12. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
13. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
14. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. Proc Conf 2016 Jun;2016:473-482 [FREE Full text] [doi: 10.18653/v1/n16-1056] [Medline: 27885364]
15. Choi E, Bahadori MT, Schuetz A, Stewart W, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. JMLR Workshop Conf Proc 2016 Aug;56:301-318 [FREE Full text] [Medline: 28286600]
16. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017 Mar 01;24(2):361-370 [FREE Full text] [doi: 10.1093/jamia/ocw112] [Medline: 27521897]
17. Zhang H, Ni W, Li J, Zhang J. Artificial intelligence-based traditional Chinese medicine assistive diagnostic system: validation study. JMIR Med Inform 2020 Jun 15;8(6):e17608 [FREE Full text] [doi: 10.2196/17608] [Medline: 32538797]
18. Zhikui C, Xin S, Jing G, Jianing Z, Peng L. Research progress in data mining-based TCM diagnoses. Chinese J Traditional Chinese Med 2020;38(12):1-9.
19. Guo Y, Ma J. Data mining technique in application of syndromes and prescriptions of Traditional Chinese medicine for diabetes (in Chinese). Med & Pharm J Chin PLA 2015;27:34-38.
20. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: bringing order to the Web. In: Proceedings of the 7th International Conference on World Wide Web. 1998 Presented at: WWW7 '98; 1998; Brisbane, Australia p. 161-172.
21. Zhang Y, Hu H, Yang T, Xie J, Shen G. Prescription recommendation algorithm of traditional Chinese medicine treatment of lung cancer based on complex network. Lishizhen Med and Materia Medica Res 2019;5:1257-1260 [FREE Full text]
22. Li W, Yang Z, Sun X. Exploration on generating Traditional Chinese medicine prescription from symptoms with an end-to-end method. In: Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing. 2019 Presented at: NLPCC '19; October 9–14, 2019; Dunhuang, China p. 486-498. [doi: 10.1007/978-3-030-32233-5_38]

23. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013 Aug;35(8):1798-1828. [doi: 10.1109/TPAMI.2013.50] [Medline: 23787338]
24. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv 2015 Aug 09 [FREE Full text] [doi: 10.48550/arXiv.1508.01991]
25. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: NAACL '16; June 12-17, 2016; San Diego, CA, USA p. 260-270. [doi: 10.18653/v1/n16-1030]
26. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017 Jul 15;33(14):i37-i48 [FREE Full text] [doi: 10.1093/bioinformatics/btx228] [Medline: 28881963]
27. Gang L, Rongqing P, Jin M, Yujie C. Entity recognition of Chinese electronic medical records based on BiLSTM-CRF network and dictionary resources. J Modern Inf 2020;40(4):3-12.
28. Li B, Kang X, Zhang H, Wang Y, Chen Y, Bai F. Named entity recognition in Chinese electronic medical records using transformer-CRF (in Chinese). Computer Engineering and Applications 2020;2020(56):153-159.
29. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the 2013 International Conference on Learning Representations. 2013 Presented at: ICLR '13; May 2-4, 2013; Scottsdale, AZ, USA.
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv 2017 Jun 12 [FREE Full text] [doi: 10.48550/arXiv.1706.03762]
31. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. In: Proceedings of the 27th Advances in Neural Information Processing Systems. 2014 Presented at: NeurIPS '14; December 8-13, 2014; Montreal, Canada. [doi: 10.1145/3422622]
32. Zhang H, Ni W, Li J, Jiang Y, Liu K, Ma Z. On standardization of basic datasets of electronic medical records in traditional Chinese medicine. Comput Methods Programs Biomed 2019 Jun;174:65-70. [doi: 10.1016/j.cmpb.2017.12.024] [Medline: 29292098]
33. Specification for drafting of health information basic dataset (WS 445-2014)S. Ministry of Health of the People's Republic of China. 2014. URL: https://wenku.baidu.com/view/4c17892d760bf78a6529647d27284b73f342365b.html [accessed 2022-05-04]
34. Specification for sharing documents of electronic medical record - part 1: medical record summary (WS/t 5001-2016)S. Ministry of Health of the People's Republic of China. 2016. URL: https://ishare.iask.sina.com.cn/f/rIcgxKYJb9.html [accessed 2022-04-21]
35. Cheung F. TCM: made in China. Nature 2011 Dec 21;480(7378):S82-S83. [doi: 10.1038/480S82a] [Medline: 22190085]
36. Xiong H. Research on the recommendation method of traditional Chinese medicine dynamic diagnosis and treatment plan based on real-world clinical data. Beijing Jiaotong University. 2020. URL: https://www.docin.com/p-2607021734.html [accessed 2022-05-04]
37. Xiaolu G. Application of deep learning in electronic health records data. Xiamen University. 2019. URL: https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFD202002&filename=1019064889.nh&uniplatform=NZKPT&v=JdsETSDw9TG7mLWRMKxEtYFZWzSE1ntYAQYDF6L2Z3Tl3ccV-citYLrD1g30mRob [accessed 2022-04-21]

Abbreviations

BiLSTM: bidirectional long short-term memory
CNN: convolutional neural network
CRF: conditional random field
EHR: electronic health record
GAN: generative adversarial network
LSTM: long short-term memory
NER: named entity recognition
NLP: natural language processing
TCM: traditional Chinese medicine

Edited by C Lovis; submitted 27.11.21; peer-reviewed by SD Boie; comments to author 08.01.22; revised version received 05.03.22; accepted 11.04.22; published 31.05.22.

Please cite as:
Zhang H, Zhang J, Ni W, Jiang Y, Liu K, Sun D, Li J
Transformer- and Generative Adversarial Network–Based Inpatient Traditional Chinese Medicine Prescription Recommendation: Development Study
JMIR Med Inform 2022;10(5):e35239
URL: https://medinform.jmir.org/2022/5/e35239/
doi:10.2196/35239
PMID:35639469

Original Paper

Characterization of Electronic Health Record Use Outside Scheduled Clinic Hours Among Primary Care Pediatricians: Retrospective Descriptive Task Analysis of Electronic Health Record Access Log Data

Selasi Attipoe[1], PhD; Jeffrey Hoffman[2,3], MD; Steve Rust[4], PhD; Yungui Huang[4], PhD; John A Barnard[3,4], MD; Sharon Schweikhart[1], MBA, PhD; Jennifer L Hefner[1,5], PhD; Daniel M Walker[5,6], PhD; Simon Linwood[4], MD

[1]Division of Health Services Management and Policy, College of Public Health, The Ohio State University, Columbus, OH, United States

[2]Division of Clinical Informatics, Nationwide Children's Hospital, Columbus, OH, United States

[3]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, United States

[4]The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

[5]Department of Family and Community Medicine, College of Medicine, The Ohio State University, Columbus, OH, United States

[6]The Center for the Advancement of Team Science, Analytics, and Systems Thinking, College of Medicine, The Ohio State University, Columbus, OH, United States

Corresponding Author:
Selasi Attipoe, PhD
Division of Health Services Management and Policy
College of Public Health
The Ohio State University
250 Cunz Hall
1841 Neil Ave
Columbus, OH, 43210
United States
Phone: 1 6144075747
Email: attipoe.1@osu.edu

Abstract

Background: Many of the benefits of electronic health records (EHRs) have not been achieved at expected levels because of a variety of unintended negative consequences such as documentation burden. Previous studies have characterized EHR use during and outside work hours, with many reporting that physicians spend considerable time on documentation-related tasks. These studies characterized EHR use during and outside work hours using clock time versus actual physician clinic schedules to define the outside work time.

Objective: This study aimed to characterize EHR work outside scheduled clinic hours among primary care pediatricians using a retrospective descriptive task analysis of EHR access log data and actual physician clinic schedules to define work time.

Methods: We conducted a retrospective, exploratory, descriptive task analysis of EHR access log data from primary care pediatricians in September 2019 at a large Midwestern pediatric health center to quantify and identify actions completed outside scheduled clinic hours. Mixed-effects statistical modeling was used to investigate the effects of age, sex, clinical full-time equivalent status, and EHR work during scheduled clinic hours on the use of EHRs outside scheduled clinic hours.

Results: Primary care pediatricians (n=56) in this study generated 1,523,872 access log data points (across 1069 physician workdays) and spent an average of 4.4 (SD 2.0) hours and 0.8 (SD 0.8) hours per physician per workday engaged in EHRs during and outside scheduled clinic hours, respectively. Approximately three-quarters of the time working in EHR during or outside scheduled clinic hours was spent reviewing data and reports. Mixed-effects regression revealed no associations of age, sex, or clinical full-time equivalent status with EHR use during or outside scheduled clinic hours.

Conclusions: For every hour primary care pediatricians spent engaged with the EHR during scheduled clinic hours, they spent approximately 10 minutes interacting with the EHR outside scheduled clinic hours. Most of their time (during and outside scheduled clinic hours) was spent reviewing data, records, and other information in EHR.

(JMIR Med Inform 2022;10(5):e34787) doi:10.2196/34787

KEYWORDS

electronic health records; access log analysis; pediatrics; primary care physicians; work outside work; work outside scheduled clinic hours

Introduction

Current research suggests that the proliferation of electronic health records (EHRs) has contributed to the increased time physicians spend interacting with computers, often at the expense of direct patient care [1-6]. Prior research has shown that physicians in the United States spend 1 to 2 additional hours completing EHR-related tasks for every hour they spend with patients [7]. Other research on this topic suggests that physicians spend approximately half their workdays on EHRs [8]. This EHR documentation burden was predicted in a systematic review published in 2005 by Canadian researchers, warning that the goal of decreased documentation time with the adoption of EHRs will likely not be realized, particularly among physicians [9].

The increased workload associated with EHR tasks has resulted in many physicians completing their EHR-related tasks during nonwork hours (eg, at night, on weekends, and during vacation time) [7,10,11]. Prior research suggests that physicians spend 90 minutes each day on EHRs outside their normal work hours. A study reported that even among physicians reporting EHR proficiency, more than half (56%) reported time spent at home on EHR-related work was *excessive* or *moderately high*, with less than one-quarter reporting sufficient time for documentation during work hours [12]. In another study, more than one-third of physicians self-reported working outside work hours, with approximately 60% of that time spent using EHRs [5]. A third study reported that of the 6 hours that clinicians spent on EHRs per weekday, 24% of this time was outside work hours [8].

Previous studies have quantified EHR work during and outside work hours [1,4-6,8,13-18] using predetermined times as their definition of work hours. Using the same approach, others have assessed the types of actions completed in EHR during these periods and the time allocated to these actions [8,15]. For instance, clerical and administrative actions (eg, documentation, order entry, billing and coding, and system security) accounted for almost half of the EHR actions (44%), and inbox management accounted for another one-quarter (24%) of that time [8].

The aim of our study is to characterize EHR work outside scheduled clinic hours among primary care pediatricians. The study design, using a retrospective descriptive task analysis of EHR access log data, extends the prior literature by identifying specific actions that are frequently completed outside work hours using physician schedules rather than fixed clock times to define outside work hours. Focusing on schedules instead of clock time allows us to produce more accurate estimates of time spent on the EHR outside of the actual scheduled clinic hours, as physician work schedules can be variable and include evenings and weekends. To our knowledge, no study thus far has used individual physician schedules to classify time spent into work and nonwork hours, which is a critical addition to the dialog and research on EHR-related documentation burden.

Methods

Setting

This study used a retrospective analysis of EHR access log data from primary care pediatricians at the Nationwide Children's Hospital (NCH), a large, free-standing US children's hospital that uses the Epic EHR (Epic Systems Corporation). All physicians who, in September 2019, generated primary care relative value units (RVUs), a measure of billable service volume and complexity, were included in the study. The use of EHR audit log data collected over a 1-month time frame is recommended because of the amount of work required to collect and clean a larger data set and the potential for shorter periods to better expose anomalies because of events such as vacations and changes in staffing [19]. Pediatricians generating non–primary care RVUs such as in inpatient or urgent care settings were omitted. All the access log data of pediatricians who met the inclusion and exclusion criteria were included in the study.

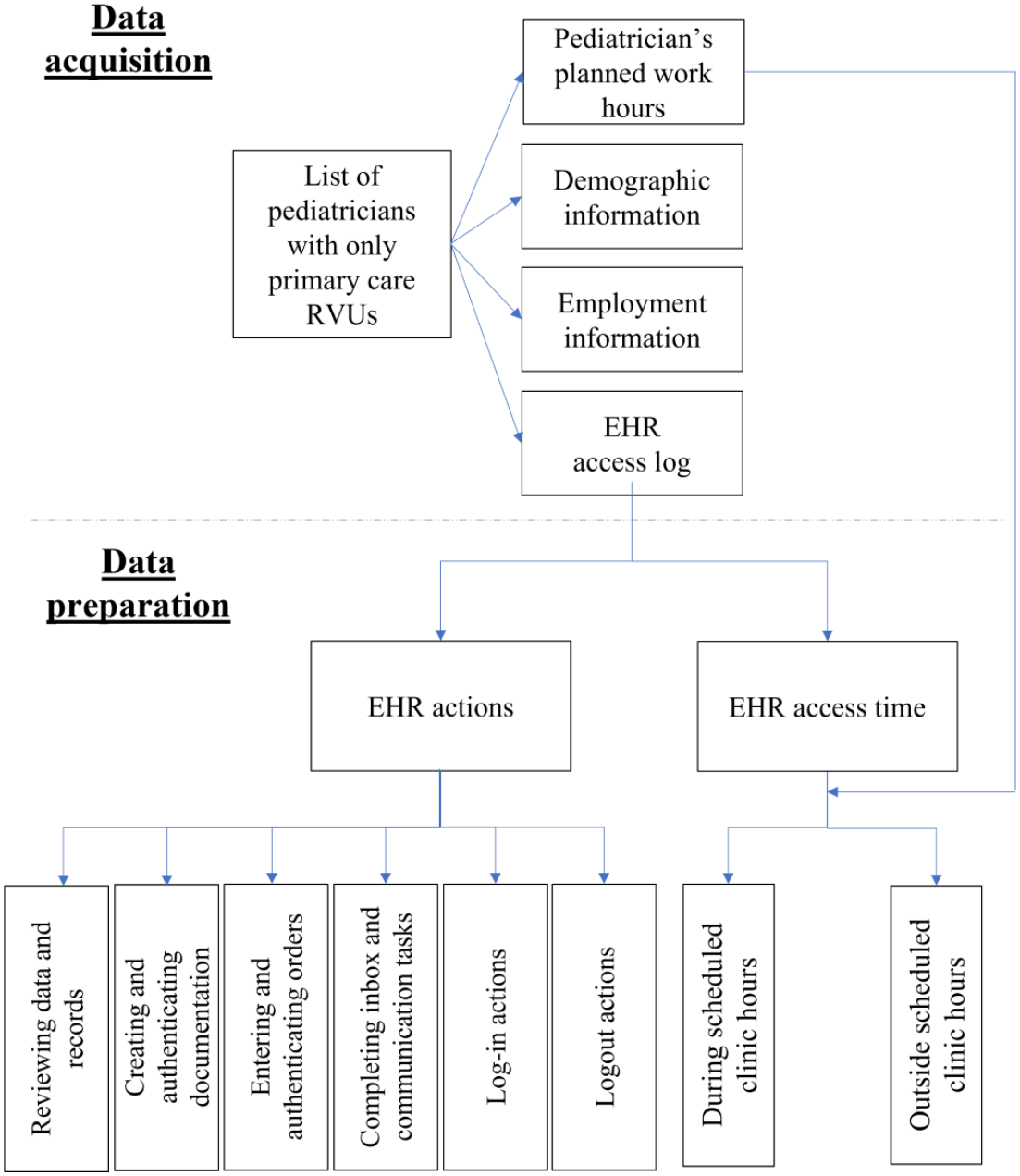
Ethics Approval

This study was approved by the institutional review boards of the NCH (protocol number IRB1800261) and Ohio State University (rotocol number 2019N0042).

Data Acquisition and Preparation

Clinical, billing, scheduling, and EHR use data were extracted from the local Epic EHR and other administrative sources into a separate database for analysis (Figure 1). The data included pediatricians' planned clinic hours, patient appointments, demographic information (eg, age and sex), employment information (length of hospital service, physicians' total workload or full-time equivalent [FTE] status, and physicians' clinical workload or clinical FTE [cFTE] status), and EHR access log entries. The *EHR access log* captures discrete time-stamped actions associated with provider navigation and use of the EHR [15,20]. It captures providers' direct interactions with the EHR system, such as log-in, logout, chart review activity, clinical documentation, and ordering actions [15,20]. Log files also record information such as the user, the time of access, the device from which the EHR was accessed, and the portion of the EHR system that was accessed [15].

Figure 1. Flow chart of data acquisition and data preparation. EHR: electronic health record; RVU: relative value unit.



The primary variables for our analysis were the *EHR actions* and *access time* extracted from the EHR access log files. *EHR actions* refer to events or movements recorded in the EHR system through mouse clicks and scrolling. These actions were grouped into 6 meaningful action categories (4 clinical and 2 general categories) using an iterative process in which the primary researcher (SA) worked with a clinical informatics physician fellow under the supervision of the NCH Chief Medical Information Officer (JH) to review various actions and associated categories. This process resulted in the identification of four clinical action categories (reviewing data and reports, creating and authenticating documentation, entering and authenticating orders, and completing inbox and communication tasks) and two general action categories (log-in and logout activities).

EHR access time (ie, duration or elapsed time) refers to the time spent in the EHR or the time spent completing actions in the EHR. Access time was estimated using a previously validated algorithm used by Arndt et al [8]. Access time was defined as the time between each activity log entry and the next log entry for a given user. The total access time was calculated for all EHR actions for each physician and then decomposed into two mutually exclusive time segments: (1) during scheduled clinic hours and (2) outside scheduled clinic hours.

EHR work during scheduled clinic hours was defined as EHR work that occurred during the period 30 minutes before to 30 minutes after scheduled patient visits for each physician each day. Similarly, *EHR work outside of scheduled clinic hours* was defined as work completed outside of the *work hours* period. A margin of 30 minutes was added to each physician's scheduled clinic hours to capture preparatory actions or closing actions for a set of consecutive patient visits. Finally, we identified and examined high users of EHR outside scheduled clinic hours to determine unique patterns of use.

Data Analysis

Descriptive task analysis was used to quantify and identify patterns of EHR work completed outside the scheduled clinic hours. All actions spanning >15 minutes were removed to omit occurrences of idle time. This cutoff was determined after careful examination of the data, sensitivity analyses, discussions with the Chief Medical Information Officer (JH), and the acknowledgment that, in practice, a single action in the EHR is typically not >15 minutes. Descriptive statistics (using demographic data) were calculated for the overall physician group. Categorical variables are reported as frequencies and percentages of the total. Continuous variables are summarized as mean and SD. The overall *EHR access time* for each physician was determined by averaging the amount of time spent during and outside the scheduled clinic hours each day across the study month. The overall time and the proportion of time spent on the *actions* completed in the EHR were examined by calculating the time *spent* per physician per workday. Administrative time (ie, time allotted within clinical schedules to complete clinical notes, inbox messages, and other administrative duties related to patient care) was calculated and reported by dividing the total number of hours of administrative time by the total number of physician workdays. The total number of administrative hours was estimated to be approximately 11% of the nominal clinical hours during the 4-week study period. The frequency (or number) and duration of EHR actions were examined to determine which actions were consistently completed outside scheduled clinic hours and whether any patterns emerged.

Regression analyses were also conducted to determine relationships between certain explanatory variables and variations in EHR use. For these analyses, the main outcome variables were the duration of EHR use both during and outside scheduled clinic hours and total EHR use. Mixed-effects statistical modeling was performed using daily and weekly aggregated data to assess the fixed effects of physician age, sex, and clinical FTE status on EHR use and estimate the magnitude of random effects because of variations among providers and temporal differences affecting all providers daily and weekly. The distributions of the outcome variables were analyzed to assess the normality assumption and determine whether a transformation was needed. All data were managed and analyzed using Microsoft Excel (version 16.0.4266) and R (version 3.5.2; R Foundation for Statistical Computing).

Results

User Statistics

There were 62 (n=14, 23% male and n=48, 77% female) pediatricians identified as working in the Division of Primary Care Pediatrics who generated primary care RVUs during September 2019, of whom 4 (6%) were excluded because they were employed on a contingency status, 1 (2%) was excluded because she had zero cFTE status, and 1 (2%) was excluded because she did not see patients during the study period. The 56 pediatricians included in the study (n=12, 21% male and n=44, 79% female) generated 1,523,872 EHR access log data points (across 1069 physician workdays). Of the 56 pediatricians, 49 (86%) used EHR outside the scheduled clinic hours. The descriptive statistics are presented in Table 1. The sample group comprised pediatricians aged 30 to 69 (mean 45.6, SD 9.9) years, with an average length of hospital service of 10.1 (SD 7.6) years (range 4 months to 33 years). The average FTE and cFTE statuses were 0.8 (SD 0.2) and 0.5 (SD 0.2), respectively.

Table 1. Descriptive statistics (N=56).

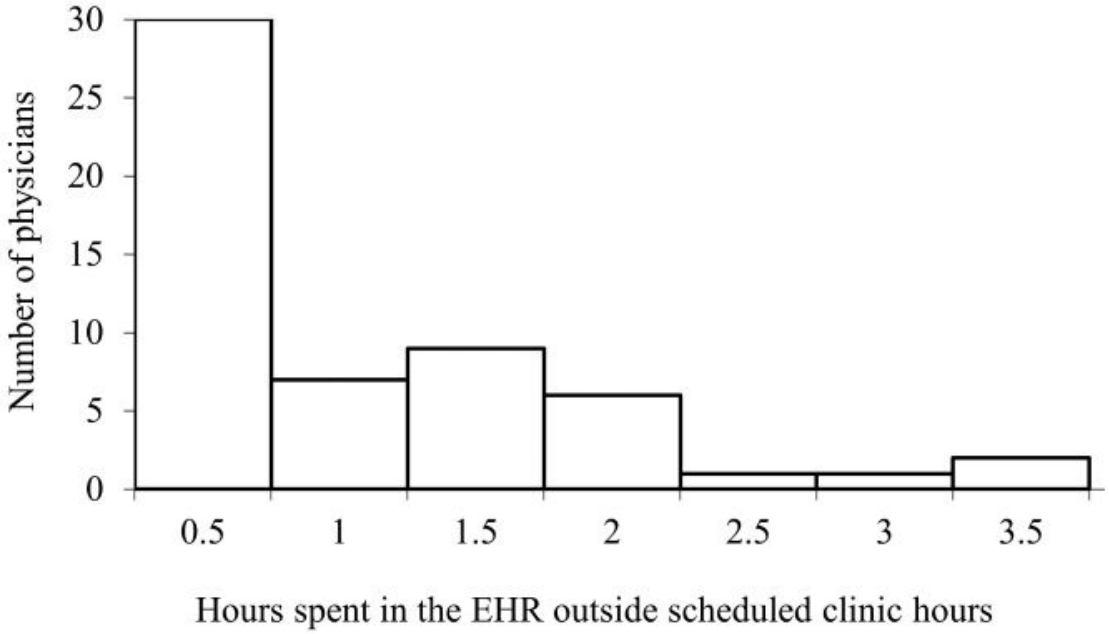
Characteristics	Values, mean (SD; range)
Age (years)	45.6 (9.9; 30-69)
Length of hospital service (years)	10.1 (7.6; 0.3-46)
Full-time equivalent status	0.8 (0.2; 0.5-1.0)
Clinical full-time equivalent status	0.5 (0.2; 0.5-0.9)
EHR[a] work during scheduled clinic hours (hours per physician per workday)	4.4 (2.0; 0.7-8.2)
EHR work outside scheduled clinic hours (hours per physician per workday)	0.8 (0.8; 0-3.2)

[a]EHR: electronic health record.

EHR Access Time

The pediatricians in this study had an average of 6 hours of scheduled work time, excluding administrative time. They spent approximately 4.4 (median 4.3) hours per workday interacting with the EHR during scheduled clinic hours and approximately 0.8 (median 0.4) hours per workday outside scheduled clinic hours. On average, the available administrative time was 0.5 hours per workday. EHR use ranged between 0.7 and 8.2 hours during scheduled clinic hours and between 0 and 3.2 hours outside of scheduled clinic hours. When physicians used the EHR outside of scheduled clinic hours, they typically did so in the evenings and on weekends. Figure 2 presents a histogram of the average time spent in the EHR by each physician outside the scheduled clinic hours.

Figure 2. Histogram of average time spent in the electronic health record (EHR) by each physician outside scheduled clinic hours.



EHR Action Categories

Overview

A total of 290 unique EHR actions were identified, and each action was classified into an EHR action category. Of the 290 EHR actions, 161 (55.5%) were classified as reviewing patient charts, 64 (22.1%) as creating and authenticating documentation, 34 (11.7%) as completing inbox and communication tasks, 19 (6.6%) as entering and authenticating orders, and 12 (4.1%) as completing log-in and logout activities.

Action Frequencies and Duration by EHR Action Categories

Table 2 presents an overview of the time spent on EHRs per physician per workday grouped by the EHR action category. Pediatricians spent approximately 73% (3.72/5.12) of their time reviewing data and reports, 7% (0.33/5.12) creating and authenticating documentation, 12% (0.60/5.12) completing inbox and communication tasks, 3% (0.13/5.12) entering and authenticating orders, and 6% (0.33/5.12) engaging in log-in and logout activities. For order entry, only 8% (0.01/0.13) of the work was completed outside scheduled clinic hours, whereas for the other 3 clinical categories, 13% (0.08/0.60) to 16% (0.59/3.72) of the work was completed outside scheduled clinic hours.

Table 2. Time spent per physician per workday by action category.

	Hours spent per physician per workday, n (%)		
	During scheduled clinic hours	Outside scheduled clinic hours	Total
Reviewing data and reports	3.13 (72)	0.59 (78)	3.72 (73)
Creating and authenticating documentation	0.28 (7)	0.05 (7)	0.33 (7)
Completing inbox and communication tasks	0.52 (12)	0.08 (11)	0.60 (12)
Entering and authenticating orders	0.12 (3)	0.01 (2)	0.13 (3)
Log-in actions	0.03 (1)	0.01 (1)	0.04 (1)
Logout actions	0.28 (6)	0.02 (2)	0.29 (6)
Total	4.35 (100)	0.76 (100)	5.12 (100)

Top 3 Most Frequent Actions by EHR Action Category

Approximately 93.1% (270/290) of EHR actions were completed outside the scheduled clinic hours per physician per workday. Of these 270 actions, the 3 most frequent specific actions completed outside scheduled clinic hours within the 4 clinical action categories accounted for 74 (27.4%) actions and 25 minutes per physician per workday (Table 3). For *chart review,* the most frequent EHR action outside scheduled clinic hours was viewing patient data, which occurred 28 times and over 13 minutes per physician per workday. This trend was similar for EHR use during scheduled clinic hours. For *documentation*, the 2 most frequent activities outside scheduled clinic hours were the use of visit documentation templates (occurring 15 times over 1 minute per physician per workday) and the signing of clinical notes (occurring 2 times over 0.4 minutes per physician

per workday). During scheduled clinic hours, the use of visit documentation templates was the most frequent activity; however, the second most frequent activity was the modification of clinical diagnoses. For *inbox and communication*, viewing inbox messages was the most frequent EHR action outside the scheduled clinic hours. This action occurred approximately 8 times over 2 minutes per physician per workday. However, during scheduled clinic hours, the most frequent EHR action in this category was the creation of inbox messages. For *order entry*, the most frequent EHR action outside scheduled clinic hours was the use of outpatient order sets, which occurred 3 times over 0.2 minutes per physician per workday. This trend was similar during the scheduled clinic hours.

Table 3. Top 3 most frequent actions completed outside scheduled clinic hours in the EHR[a] per physician per workday by EHR action category.

	Frequency per physician per workday	Total minutes spent per physician per workday
Reviewing data and reports		
Patient data viewed	28	12.7
Encounter data viewed	4	3.5
Clinical notes viewed	4	3.2
Creating and authenticating documentation		
Visit documentation template used	15	1.1
Clinical note signed	2	0.4
Encounter diagnoses entered	2	0.3
Completing inbox and communication tasks		
Inbox message viewed	8	2.2
Inbox message created	3	0.7
Inbox folder loaded	3	0.7
Entering and authenticating orders		
Outpatient order sets used	3	0.2
Order list changed	1	0.2
Length of stay entered	1	0.2
Total	74	25.4

[a]EHR: electronic health record.

High Outside Scheduled Clinic Hours EHR Users

EHR use by physicians who spent >1.5 hours per workday outside scheduled clinic hours (10/56, 18%) was further examined to determine if there were additional insights that could be gained from pediatricians who use the EHR more outside scheduled clinic hours. Together, these physicians generated a total of 212 physician workdays, spent an average of 2.2 hours per physician per workday in the EHR outside scheduled clinic hours, and exhibited similar trends (in terms of the most frequent activities completed in the EHR) to those of the entire group.

Factors Associated With EHR Use

Mixed-effects models revealed no significant associations of age, sex, and cFTE status with EHR use during or outside scheduled clinic hours (Table 4).

Table 4. Mixed regression models.

Models	EHR[a] work during scheduled clinic hours	EHR work outside scheduled clinic hours	Total EHR use
Fixed effects, coefficients (SE)			
EHR work during scheduled clinic hours (minutes)	N/A[b]	−0.18 (0.02)	N/A
Age (years)	−0.06 (0.02)	−0.004 (0.01)	−0.05 (0.02)
Gender	0.86 (0.50)	−0.80 (0.29)	−0.14 (0.43)
cFTE[c] status	3.63 (0.92)	0.67 (0.54)	3.63 (0.79)
Constant	3.00 (1.34)	2.23 (0.76)	4.67 (1.18)
Random effects, variance (SD)			
Day	3.45 (1.86)	0.06 (0.24)	3.74 (1.93)
Provider	1.91 (1.38)	0.63 (0.80)	1. 35 (1.61)
Model fitness (R^2; %)			
Fixed effects	10.0	9.9	7.8
Random effects	41.4	53.4	41.5
Total	51.4	63.7	49.3

[a]EHR: electronic health record.

[b]N/A: not applicable.

[c]cFTE: clinical full-time equivalent.

Discussion

Principal Findings

In this study, we quantified and characterized EHR work outside scheduled clinic hours and found that pediatricians spent approximately 0.8 hours per physician per workday completing work in the EHR outside of scheduled clinic hours. The time spent using the EHR outside scheduled clinic hours accounted for approximately 15% of the total daily EHR time (ie, 5 hours per physician per workday). Specifically, outside scheduled clinic hours (ie, 0.76 hours per physician per workday), pediatricians spent 78% of their time (ie, 0.59 hours per physician per workday) reviewing data and reports, 11% (ie, 0.08 hours per physician per workday) completing inbox and communication tasks, 8% (ie, 0.06 hours per physician per workday) documenting and completing orders, and 3% (ie, 0.03 hours per physician per workday) engaging in log-in and logout activities. This distribution across action categories was similar to the distribution of actions during scheduled clinic hours.

Comparison With Prior Work

The *proportion of total time spent in the EHR outside work hours* in this study (0.76/5.12, 15%) was lower than that reported by Arndt et al (24%) [8], Rotenstein et al (25% and 26%) [14,18], and Holmgren et al (30%) [13] and higher than that reported by Overhage and Johnson (12%) [21] and Holmgren et al (13%) [17]. Each of these alternative estimates characterized EHR use outside work hours using a predefined clock time, whereas this study used actual physician schedules. The methodology used in this study is arguably superior because of the granular level of detail used to classify time as during or outside work hours. We used scheduled patient visits (and physician schedules by extension) to define the EHR work outside work hours for each physician. We also included 30 minutes before and after each scheduled clinic time to capture preparatory and closing actions. Thus, we are confident that our time segment classifications truly reflect whether a physician was actively seeing patients or completing related tasks. Using clock time to define work versus after-work time might not always capture exactly when a physician starts and ends their actual workday.

The daily *time spent in the EHR* reported in this study (ie, 5 hours) is comparable with current estimates in the literature, which range from 1.5 to 5 hours [1,6,8,13,15,18,22]. With respect to the time spent per action category, most of the pediatricians' time was spent *reviewing data and reports* both during and outside scheduled clinic hours. Only a small fraction of their time was spent completing *documentation and order entry actions.* This finding differs from reports in the literature and anecdotal evidence that indicate physicians spend most of their time completing documentation-related activities, particularly outside work hours [1-6,8,13,15,23,24]. For instance, the study by Overhage and Johnson [21] found that among their sample of pediatricians practicing in US-based ambulatory practices, documentation accounted for 31% of EHR use time, and chart review accounted for another 31% of EHR use time. The study by Arndt et al [8] found that nonteaching ambulatory physicians spent 44% of the total EHR use time engaged in clerical and administrative tasks (eg, documentation, order entry, billing and coding, and system security). Another study by Tai-Seale et al [15] found that primary care physicians spent 51% of their time completing EHR work, and 34% of this portion was spent on progress notes. A more recent study by Holmgren et al [13] found that ambulatory clinicians in the United States spent 67% of their EHR use time completing notes and orders.

One of the reasons why documentation time estimates in this study were lower than those commonly reported in the literature may be the differences in the categorization of EHR actions. In the abovementioned studies, the time spent viewing patient data during the process of writing a progress note may not have been distinguished from the total time spent on the note (ie, from the time the note was opened until it was finally signed). In this study, raw access log data were used to categorize each EHR action into one of the EHR action categories. No meanings were inferred—all viewing actions were categorized under *reviewing data and reports*, whereas all data entry actions were categorized as *documentation*. The level of granularity and objectivity used in this study ensures the robustness of our categorization and estimates. Another reason for the relatively lower documentation time estimates may be attributed to the extensive use of documentation templates with quick selection options and prepopulated data, as well as the extensive use of outpatient order sets for good childcare and common presenting complaints at this hospital. These practices may contribute to the reduced time spent on the EHR on documentation activities.

In addition, we found that pediatricians spend approximately 10% of their time completing *inbox and communication actions* both during and outside scheduled clinic hours. This estimate is lower than that reported in prior studies. The study by Holmgren et al [13] found that inbox activities accounted for approximately 14% of EHR work, whereas Arndt et al [8] and Tai-Seale et al [15] reported 24% and 22%, respectively. Interestingly, in this study, the loading and viewing of inbox messages were the most frequent and longest EHR actions outside of scheduled clinic hours in the *communication* category; however, during scheduled clinic hours, the most frequent and longest EHR action was the creation of messages. Perhaps physicians spend time checking their messages outside scheduled clinic hours to stay abreast with current patient needs but wait to respond to these messages during their scheduled clinic hours. A second study by Tai-Seale et al [25] found that receiving an excessive amount of system-generated inbox messages was associated with a higher probability of burnout and intention to reduce clinical work time, suggesting that this aspect of EHR work can have considerable effects on a physician's well-being. On the other hand, a more recent study by Melnick et al [26] suggested that EHR inbox management was associated with physician departure—less time spent on EHRs was associated with physician departure. Although their finding was counterintuitive, they proposed that tracking EHR metrics could potentially identify physicians at a high risk of departure [26].

Factors Associated With EHR Use

This study found no association of age, sex, or cFTE with EHR use. This finding is in contrast to previous findings from this research group [27]. In our previous work, we found that female physicians spend more time than male physicians using the EHR during work hours but not outside work hours. Provider-to-provider variation was the largest and most dominant source of variation in EHR use outside work hours, accounting for 52% of the total variance. However, in that study, EHR work outside work hours was defined using clock time, whereas our approach in this study of using actual physician schedules may have produced more accurate estimations, which could have eliminated bias in our prior models that accounted for the observed differences.

EHR Access Log Data Use in Research

The use of EHR access logs is valid for assessing EHR actions as there is consistency between the direct observation findings, physician self-reported EHR work outside work hours, and EHR system event log data [8,28]. Although EHR access log data are highly complex, often uncharacterized, and require powerful statistical software and technical skills for processing and analysis, understanding and using raw EHR data could serve as an external validation of EHR vendor–supplied metrics. This validation is important as many researchers and hospital administrators use vendor-supplied data to explore research questions because of their ready availability and ease of use. However, the proprietary algorithms used by EHR vendors (eg, Epic's Signal and Cerner's LightsOn) use a *black box* methodology, wherein the actual composition of each metric is unknown. This study remedies this limitation by exposing and using raw access log data to produce more meaningful metrics and analyses.

At present, there are no agreed-upon standards for categorizing EHR actions. A standard categorization scheme for EHR actions will help provide a common language to facilitate and promote clear communication in this nascent research space. Upon close review of action categories used in the abovementioned studies, other studies in the scientific literature, and this study, the following *conceptual* categorization scheme of clinical EHR actions seems adequate as a foundation on which to further build: data review, data entry, data transmission, and other (Table 5). Rather than creating new action categories with each new research study, we propose building on the aforementioned categories, as this classification scheme is both clear and clinically meaningful. As it relates to this study, the proposed conceptual classification scheme aligns well with the categories used in this study in that data review aligns with *reviewing data and records*, data entry aligns with *creating and authenticating documentation* and *entering and authenticating orders*, data transmission aligns with *completing inbox and communication tasks*, and *other* aligns with *log-in and logout activities*.

The set of EHR action categories used by Zheng et al [29] (ie, reading, entering, printing, processing, log-in, and logout) is arguably one of the clearest among the available classifications used in the literature as it objectively categorizes the action without assigning any meaning—for instance, reading versus chart review. Perhaps, this strength is also the reason researchers refrain from using it; that is, the categories lack clinical meaning. The action categories used by Arndt et al [8] (ie, medical care, clerical, and inbox) have the opposite issue: they have clinical meaning but are somewhat ambiguous. For instance, some of the actions in the category of *medical care* could also be seen as clerical tasks. The categories used by Holmgren et al [13] closely align with those used in this study, are clear, and have clinical meaning.

Table 5. Proposed conceptual EHR[a] action categorization scheme.

Conceptual EHR action categorization scheme	Action categories used in this study	Action categories used by Holmgren et al [13]	Action categories used by Arndt et al [8]	Action categories used by Zheng et al [29]
Data review—information review, retrieval, or gathering activities	Reviewing data and reports	Clinical review	Medical care	Reading
Data entry—information entry or recording activities	Creating and authenticating documentation; entering and authenticating orders	Notes; orders	Clerical	Entering
Data transmission—information transmission activities	Inbox and communication tasks	In-basket messages	Inbox	Printing
Other—other nonclinical activities	Log-in and logout activities	N/A[b]	N/A	Log-in, logout, and processing

[a]EHR: electronic health record.

[b]N/A: not applicable.

Strengths and Limitations

A major strength of this study is the level of objectivity and granularity used to define time segments and action categories. Time segments were defined using actual scheduled patient visits to construct the physician workday schedules. These schedules were validated against the planned physician schedules. To the best of our knowledge, no study has used actual schedules to define EHR outside of work hours. In addition, action categories were defined using clear and objective criteria. Such a concise categorization of EHR work outside work hours and EHR action categories facilitates more accurate estimations. However, there are a few limitations to this study.

First, we were unable to parse chart review actions associated with other action categories. Several chart review actions are associated with other action categories. For instance, documentation-related actions are usually associated with chart review actions, and the methodology used in this study did not capture these nuanced associations. For example, if a physician viewed previous clinical notes while writing their own clinical note for that encounter, this action was classified as *reviewing data and reports*; however, to the physician, this viewing action might be more cognitively associated with documentation. This may explain why our estimates for the *reviewing data and reports* category were relatively high. This explanation also addresses why we found that physicians in this study spent only a small fraction of their time completing documentation and order entry actions, which was lower than the estimates in the literature and anecdotal evidence. The current scientific literature suggests that physicians spend a considerable amount of time outside work hours completing documentation-related activities [8], although the EHR is purported to contribute to more efficient use of physicians' time.

In addition, this study did not have (and therefore did not include) work RVU (wRVU) as a factor in the regression analysis. wRVU is a key factor for understanding EHR work. Typically, wRVU indicates the volume and intensity of medical services provided; thus, the higher the wRVU, the more likely it is for a physician to spend time with the EHR. The absence of wRVU in the regression models may be the reason they did not generate statistically significant associations. However, the findings are important as they provide a general characterization of EHR use by pediatricians at this institution.

Finally, the study sample size was limited to 1 calendar month of EHR activity data for a single practice, setting, type of provider, and commercial EHR system (ie, academic primary care pediatricians at NCH using the Epic system), thus limiting the generalizability of the study findings. For instance, our study findings may not be generalizable to other specialties, including primary care specialties for adults, nor are they likely generalizable to subspecialty academic practices—they are most relevant to the academic pediatric practice. Furthermore, EHR interfaces are often modified according to the needs of each provider system [22]. Thus, the reported EHR use statistics may not be generalizable to other institutions and provider groups. On the other hand, Epic's EHR is the most widely used EHR in the United States and includes use metrics [30], making our findings widely comparable with other institutions and provider groups. Furthermore, the pediatric population is an important one, and the sample group (ie, primary care physicians) helps reduce the technical complexity of studying work during and outside work hours.

Implications and Future Research

Primary care pediatricians care for many children during half-day sessions (often simultaneously), work with nurses and other support staff, and interact with patients at multiple points in their daily workflow [28]. In addition, they spend considerable time on EHRs during and outside work hours to document and provide care. Thus, there is a need to improve physician-computer interactions by streamlining EHR workflows [22]. These improvements will likely need to be customized so that they are relevant to the specific type of practice: general pediatrics, subspecialty pediatrics, and many variations of adult practices. To identify interventions to improve EHR design and use, physicians' EHR actions must be properly characterized to better understand their various activities and use patterns [22]. By identifying specific EHR actions that consistently dominate computer use across multiple providers, more targeted, data-driven approaches could be developed to improve physician-computer interactions [22]. This implication reinforces the need to validate proprietary algorithms and metrics generated by EHR vendors, as many researchers and hospital

administrators rely on these metrics (vs computing them from raw EHR log data) for clinical, research, and policy purposes. This is understandable, given the tremendous complexity and resource requirements for working with and processing raw EHR access log data.

Contrary to prior research and anecdotal evidence, our analysis found that pediatricians spend a moderate amount of time on EHRs outside of scheduled clinic hours and relatively less time completing documentation-related tasks. As described previously, this hospital uses documentation templates extensively, which potentially helps reduce documentation time in the EHR. Other medical facilities may consider adopting such usability features to reduce the documentation burden among providers. With the issue of EHR documentation burden being prevalent among physicians and contributing to burnout among this group [31-35], opportunities to reduce the burden may help enhance physician well-being.

There are many opportunities for future research in this area, including standardizing vendor-derived EHR data descriptions in a way that is clinically relevant and important [26], validating the use of EHR access log data across different settings, exploring the relationship between EHR action frequency and EHR action duration, examining the contribution of EHR use outside work hours to physician well-being, determining overestimation and underestimation margins of estimates, and developing a taxonomy of EHR use to further promote consistency and valid comparisons across organizations and research studies. Such research will help provide additional insights into EHR workflow issues and the effect of EHR work on physician well-being. Furthermore, researchers in this field should strive to set standards [36,37], as we have proposed above. Accepted standards, for instance, on how to calculate work outside work hours and categorize EHR actions, will help facilitate research in this space.

Conclusions

In this study, we used EHR access log data to identify actions typically completed outside scheduled clinic hours and the pattern of this EHR work. This study fills a gap in the literature by quantifying the use of EHR outside of scheduled clinic hours using actual scheduled patient visits rather than planned physician schedules or predefined clock times as a proxy. The findings from this study suggest that primary care pediatricians spend more than one-tenth of their EHR use time outside of scheduled clinic hours and that approximately three-quarters of this time is spent reviewing data and reports, whereas negligible time is spent completing orders. Further studies are needed to explore EHR use patterns by physicians and the reasons for these patterns to help improve EHR work and workflow. Qualitative and mixed methods research studies will be instrumental in gaining insights into these patterns.

Acknowledgments

The authors would like to thank the following individuals: Dr Alex Kemper; Dr Dane Snyder; Rajesh Ganta; Samuel Yang; Richard Hoyt; Eric Yu; Macy Rees; and Elaine Damo and her Data Resource Center team for providing access to the data, extracting data, and/or helping us understand the data for this study. The opinions and assertions expressed herein are those of the author and should not be construed as reflecting those of the Ohio State University or the Nationwide Children's Hospital.

Conflicts of Interest

None declared.

References

1. Fletcher KE, Visotcky AM, Slagle JM, Tarima S, Weinger MB, Schapira MM. The composition of intern work while on call. J Gen Intern Med 2012 Nov;27(11):1432-1437 [FREE Full text] [doi: 10.1007/s11606-012-2120-7] [Medline: 22865015]
2. Oxentenko AS, Manohar CU, McCoy CP, Bighorse WK, McDonald FS, Kolars JC, et al. Internal medicine residents' computer use in the inpatient setting. J Grad Med Educ 2012 Dec;4(4):529-532 [FREE Full text] [doi: 10.4300/JGME-D-12-00026.1] [Medline: 24294435]
3. Carayon P, Wetterneck TB, Alyousef B, Brown RL, Cartmill RS, McGuire K, et al. Impact of electronic health record technology on the work and workflow of physicians in the intensive care unit. Int J Med Inform 2015 Aug;84(8):578-594 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.04.002] [Medline: 25910685]
4. Oxentenko AS, West CP, Popkave C, Weinberger SE, Kolars JC. Time spent on clinical documentation: a survey of internal medicine residents and program directors. Arch Intern Med 2010 Feb 22;170(4):377-380. [doi: 10.1001/archinternmed.2009.534] [Medline: 20177042]
5. 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]
6. Cox ML, Farjat AE, Risoli TJ, Peskoe S, Goldstein BA, Turner DA, et al. Documenting or operating: where is time spent in general surgery residency? J Surg Educ 2018 Nov;75(6):e97-e106. [doi: 10.1016/j.jsurg.2018.10.010] [Medline: 30522828]
7. Wright AA, Katz IT. Beyond burnout - redesigning care to restore meaning and sanity for physicians. N Engl J Med 2018 Jan 25;378(4):309-311. [doi: 10.1056/NEJMp1716845] [Medline: 29365301]

8. 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]
9. Poissant L, Pereira J, Tamblyn R, Kawasumi Y. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. J Am Med Inform Assoc 2005;12(5):505-516 [FREE Full text] [doi: 10.1197/jamia.M1700] [Medline: 15905487]
10. Gregory ME, Russo E, Singh H. Electronic health record alert-related workload as a predictor of burnout in primary care providers. Appl Clin Inform 2017 Jul 05;8(3):686-697 [FREE Full text] [doi: 10.4338/ACI-2017-01-RA-0003] [Medline: 28678892]
11. Miyasaki JM, Rheaume C, Gulya L, Ellenstein A, Schwarz HB, Vidic TR, et al. Qualitative study of burnout, career satisfaction, and well-being among US neurologists in 2016. Neurology 2017 Oct 17;89(16):1730-1738. [doi: 10.1212/WNL.0000000000004526] [Medline: 28931640]
12. Kroth PJ, Morioka-Douglas N, Veres S, Pollock K, Babbott S, Poplau S, et al. The electronic elephant in the room: physicians and the electronic health record. JAMIA Open 2018 Jul;1(1):49-56 [FREE Full text] [doi: 10.1093/jamiaopen/ooy016] [Medline: 31093606]
13. 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]
14. 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]
15. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff (Millwood) 2017 Apr 01;36(4):655-662 [FREE Full text] [doi: 10.1377/hlthaff.2016.0811] [Medline: 28373331]
16. Young RA, Burge SK, Kumar KA, Wilson JM, Ortiz DF. A time-motion study of primary care physicians' work in the electronic health record era. Fam Med 2018 Feb;50(2):91-99 [FREE Full text] [doi: 10.22454/FamMed.2018.184803] [Medline: 29432623]
17. Holmgren AJ, Lindeman B, Ford EW. Resident physician experience and duration of electronic health record use. Appl Clin Inform 2021 Aug;12(4):721-728. [doi: 10.1055/s-0041-1732403] [Medline: 34348409]
18. Rotenstein LS, Holmgren AJ, Downing NL, Longhurst CA, Bates DW. Differences in clinician electronic health record use across adult and pediatric primary care specialties. JAMA Netw Open 2021 Jul 01;4(7):e2116375 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.16375] [Medline: 34241631]
19. 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]
20. Adler-Milstein J, Huckman RS. The impact of electronic health record use on physician productivity. Am J Manag Care 2013 Nov;19(10 Spec No):SP345-SP352 [FREE Full text] [Medline: 24511889]
21. Overhage JM, Johnson KB. Pediatrician electronic health record time use for outpatient encounters. Pediatrics 2020 Dec;146(6):e20194017. [doi: 10.1542/peds.2019-4017] [Medline: 33139456]
22. 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;14(2):e0205379 [FREE Full text] [doi: 10.1371/journal.pone.0205379] [Medline: 30726208]
23. Christino MA, Matson AP, Fischer SA, Reinert SE, Digiovanni CW, Fadale PD. Paperwork versus patient care: a nationwide survey of residents' perceptions of clinical documentation requirements and patient care. J Grad Med Educ 2013 Dec;5(4):600-604 [FREE Full text] [doi: 10.4300/JGME-D-12-00377.1] [Medline: 24455008]
24. Friedberg MW, Chen PG, Van Busum KR, Aunon F, Pham C, Caloyeras J, et al. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand Health Q 2014 Dec 1;3(4):1 [FREE Full text] [Medline: 28083306]
25. Tai-Seale M, Dillon EC, Yang Y, Nordgren R, Steinberg RL, Nauenberg T, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019 Jul;38(7):1073-1078. [doi: 10.1377/hlthaff.2018.05509] [Medline: 31260371]
26. Melnick ER, Fong A, Nath B, Williams B, Ratwani RM, Goldstein R, et al. Analysis of electronic health record use and clinical productivity and their association with physician turnover. JAMA Netw Open 2021 Oct 01;4(10):e2128790 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.28790] [Medline: 34636911]
27. Attipoe S, Huang Y, Schweikhart S, Rust S, Hoffman J, Lin S. Factors associated with electronic health record usage among primary care physicians after hours: retrospective cohort study. JMIR Hum Factors 2019 Sep 30;6(3):e13779 [FREE Full text] [doi: 10.2196/13779] [Medline: 31573912]

28. Hribar MR, Read-Brown S, Goldstein IH, Reznick LG, Lombardi L, Parikh M, et al. Secondary use of electronic health record data for clinical workflow analysis. J Am Med Inform Assoc 2018 Jan 01;25(1):40-46 [FREE Full text] [doi: 10.1093/jamia/ocx098] [Medline: 29036581]
29. Zheng K, Ciemins EL, Lanham HJ, Lindberg C, Man D. Examining the relationship between health it and ambulatory care workflow redesign - final report. Agency for Healthcare Research and Quality. URL: https://psnet.ahrq.gov/issue/examining-relationship-between-health-it-and-ambulatory-care-workflow-redesign [accessed 2022-04-18]
30. Baxter SL, Apathy NC, Cross DA, Sinsky C, Hribar MR. Measures of electronic health record use in outpatient settings across vendors. J Am Med Inform Assoc 2021 Apr 23;28(5):955-959 [FREE Full text] [doi: 10.1093/jamia/ocaa266] [Medline: 33211862]
31. Collier R. Electronic health records contributing to physician burnout. CMAJ 2017 Nov 13;189(45):E1405-E1406 [FREE Full text] [doi: 10.1503/cmaj.109-5522] [Medline: 29133547]
32. 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]
33. DiAngi YT, Longhurst CA, Payne TH. Taming the EHR(electronic health record) - there is hope. J Fam Med 2016;3(6):1072 [FREE Full text] [Medline: 27830215]
34. Harris DA, Haskell J, Cooper E, Crouse N, Gardner R. Estimating the association between burnout and electronic health record-related stress among advanced practice registered nurses. Appl Nurs Res 2018 Oct;43:36-41. [doi: 10.1016/j.apnr.2018.06.014] [Medline: 30220361]
35. 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]
36. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. J Am Med Inform Assoc 2020 Mar 01;27(3):480-490 [FREE Full text] [doi: 10.1093/jamia/ocz196] [Medline: 31750912]
37. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]

Abbreviations

cFTE: clinical full-time equivalent
EHR: electronic health record
FTE: full-time equivalent
NCH: Nationwide Children's Hospital
RVU: relative value unit
wRVU: work relative value unit

Edited by C Lovis; submitted 10.12.21; peer-reviewed by B Arndt, O Kanste; comments to author 08.01.22; revised version received 01.03.22; accepted 27.03.22; published 12.05.22.

Please cite as:
Attipoe S, Hoffman J, Rust S, Huang Y, Barnard JA, Schweikhart S, Hefner JL, Walker DM, Linwood S
Characterization of Electronic Health Record Use Outside Scheduled Clinic Hours Among Primary Care Pediatricians: Retrospective Descriptive Task Analysis of Electronic Health Record Access Log Data
JMIR Med Inform 2022;10(5):e34787
URL: https://medinform.jmir.org/2022/5/e34787
doi:10.2196/34787
PMID:35551055

Original Paper

Integrated Health Record Viewers and Reduction in Duplicate Medical Imaging: Retrospective Observational Analysis

Yingzhe Yuan[1,2], MPH; Megan Price[2], MS; David F Schmidt[3,4], MD; Merry Ward[3], PhD; Jonathan Nebeker[3,5], MS, MD; Steven Pizer[1,2], PhD

[1]Department of Health Law, Policy & Management, School of Public Health, Boston University, Boston, MA, United States

[2]Partnered Evidence-Based Policy Resource Center, Boston Veterans Affairs Healthcare System, Boston, MA, United States

[3]Veterans Health Administration Office of Health Informatics, Department of Veterans Affairs, Washington DC, DC, United States

[4]School of Medicine, Oregon Health and Science University, Portland, OR, United States

[5]School of Medicine, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Yingzhe Yuan, MPH
Partnered Evidence-Based Policy Resource Center
Boston Veterans Affairs Healthcare System
Building 9, 4th Floor
150 S Huntington Ave
Boston, MA, 02130
United States
Phone: 1 857 364 5479
Email: yuanyz@bu.edu

Abstract

Background: Health information exchange and multiplatform health record viewers support more informed medical decisions, improve quality of care, and reduce the risk of adverse outcomes due to fragmentation and discontinuity in care during transition of care. An example of a multiplatform health record viewer is the VA/DoD Joint Longitudinal Viewer (JLV), which supports the Department of Veterans Affairs (VA) and Department of Defense (DoD) health care providers with read-only access to patient medical records integrated from multiple sources. JLV is intended to support more informed medical decisions such as reducing duplicate medical imaging when previous image study results may meet current clinical needs.

Objective: We estimated the impact of provider usage of JLV on duplicate imaging for service members transitioning from the DoD to the VA health care system.

Methods: We conducted a retrospective cross-sectional study in fiscal year 2018 to examine the relationship between providers' use of JLV and the likelihood of ordering duplicate images. Our sample included recently separated service members who had a VA primary care visit in fiscal year 2018 within 90 days of a DoD imaging study. Patients who received at least one imaging study at VA within 90 days of a DoD imaging study of the same imaging mode and on the same body part are considered to have received potentially duplicate imaging studies. We use a logistic regression model with "JLV provider" (providers with 1 or more JLV audits in the prior 6 months) as the independent variable to estimate the relationship between JLV use and ordering of duplicate images. Control variables included provider image ordering rates in the prior 6 months, provider type, patient demographics (age, race, gender), and clinical characteristics (Elixhauser comorbidity score).

Results: Providers known to utilize JLV in the prior 6 months order fewer duplicate images relative to providers not utilizing JLV for similar visits over time (odds ratio 0.44, 95% CI 0.24-0.78; *P*=.005). This effect is robust across multiple specifications of linear and logistic regression models. The provider's practice pattern of ordering image studies and the patient's health status are powerful confounders.

Conclusions: This study provides evidence that adoption of a longitudinal viewer of health records from multiple electronic health record systems is associated with a reduced likelihood of ordering duplicate images. Investments in health information exchange systems may be effective ways to improve the quality of care and reduce adverse outcomes for patients experiencing fragmentation and discontinuity of care.

(JMIR Med Inform 2022;10(5):e32168) doi:10.2196/32168

KEYWORDS

health informatics; duplicate medical imaging; health record viewer; health care system; health care; health records; electronic health records; health information exchange

Introduction

Health information exchange (HIE) allows health care providers and patients to access and share patient-level electronic health information between different health care settings [1-3]. When health information such as radiology reports, laboratory results, and drug allergy history is shared, HIE helps ensure the safety of patients and improve clinic efficiency [4]. Adoption of HIE has the potential to address the Institute of Medicine's quality aims [5] and produce substantial financial value [6]. Previous research linked the use of electronic health record viewers or HIE participation to improved health care quality measures such as higher patient satisfaction or lower readmission and duplicate diagnostic imaging study rates. A recent study by Legler et al [7] demonstrated that providers' early adoption of a longitudinal health record viewer was related to patients more likely reporting that providers were knowledgeable of their medical history. In another recent work, Chen et al [8] found that hospitals' participation in HIE was associated with a reduction in 30-day readmission rates in Florida. Bailey et al [9] found that use of HIE was associated with a decreased probability of ordering repeated diagnostic imaging in the emergency evaluation of back pain. For patients transitioning between health care systems, fragmentation and discontinuity in care increase the risk of adverse health outcomes [10]. Providers' access to patient-level medical records from multiple health care settings may support more informed medical decisions, improve quality of care, enhance care coordination, and reduce risks of adverse outcomes due to fragmentation.

Minimizing orders for unnecessary duplicate medical image studies is important for improving health care efficiency, reducing unnecessary time burdens on patients, and attenuating adverse health outcomes caused by excessive medical radiation, such as increased risk of cancer [10-13]. However, there is little evidence regarding the impact of HIE on duplicate imaging. Vest et al [14] found that the use of an HIE system to access previous patient information was associated with a reduction in repeated medical imaging, but the study was limited by its setting in 11 counties in New York and was unable to adjust for potential confounders at the provider level. In this paper, we estimate the impact of provider usage of integrated health record viewers on the ordering of duplicate imaging for patients receiving health care in multiple settings.

An example of integrated health record viewers is the Joint Longitudinal Viewer (JLV), formerly known as the Joint Legacy Viewer (version 2.2). As a web-based graphical user interface, JLV supports the Department of Veterans Affairs (VA) and Department of Defense (DoD) health care providers with an integrated, read-only view of health data from the VA and DoD systems as well as VA community partners [7]. Released on October 1, 2014, JLV has been used by an increasing number of providers to view noncomputable patient-level health information such as vital signs, physician notes, medications, allergy, immunization, and radiology records [15]. The integrated viewer allows providers to access a complete set of the patient's previous medical images and therefore has the potential to reduce the frequency of duplicate medical image studies.

Methods

Study Design

We conducted a retrospective cross-sectional study in fiscal year 2018 to examine the relationship between provider use of JLV and the ordering of potentially duplicate image studies. The analysis compared duplicate imaging ordered by JLV-using and non–JLV-using providers of VA outpatient primary care visits in fiscal year 2018 for recently separated service members. We conducted the study for VA quality improvement and program evaluation purposes, and therefore, the study was exempt from Institutional Review Board review.

Participants and Setting

Recently separated service members who had at least one VA primary care visit in fiscal year 2018 within 90 days of an imaging study conducted at DoD were eligible to be included in the sample. We excluded VA primary care visits that were compensation and pension exams or not provided by physicians, physician assistants, or nurse practitioners. We also excluded DoD imaging studies if the primary diagnosis was cancer because duplicate diagnostic images were likely to be clinically appropriate and recommended by providers for patients with cancer. Patients who received at least one imaging study at VA within 90 days of a DoD imaging study using the same imaging mode and on the same body part were considered to have received potentially duplicate imaging studies [11].

Measures

VA clinic stop codes (322, 323, and 350) were used to identify outpatient primary care visits. Compensation and Pension exams were identified using the secondary stop code (450) and the appointment type (Compensation and Pension) and were excluded from the VA primary care visits. Audit logs acquired from the JLV system were assessed to determine a provider's JLV utilization during a specific VA primary care visit and the provider's JLV utilization history over the 6 months prior to the visit.

The independent variable "JLV encounter" indicated whether a JLV audit was linked to the patient on the primary care visit date. The independent variable "JLV provider" indicated whether the provider had 1 or more JLV audits in the 6 months prior to the visit date. Endogeneity is likely to be a problem in the estimation of the association between "JLV encounter" and duplicate imaging because unobserved confounders such as patient complexity are related to both JLV use and ordering duplicate image studies during the primary care visit. We estimated the direct (proxy) relationship between "JLV provider" and duplicate imaging to deal with the potential endogeneity problem. We also used a 2-stage statistical model by using JLV

providers as an instrumental variable to estimate the causal relationship between JLV encounter and duplicate imaging. The categorization of JLV providers was based on the provider's prior interactions with other patients and indicated the provider's propensity to view health records through JLV. Thus, this variable was independent of the observed and unobserved characteristics of the patient under study. This is especially true in the VA setting where patients are arbitrarily assigned to primary care providers. Current Procedural Terminology codes indicating imaging procedures were categorized by mode and body part to compare VA and DoD imaging records and identify potential duplicate images. Following Vest et al [14], the dependent variable was coded as "duplicate image" if an imaging study ordered during the VA primary care visit was of the same mode and the same body part as a DoD imaging study for the patient within 90 days prior to the VA visit date. Covariates included the provider's rate of ordering images during previous primary care visits with other patients over the 6 months prior to the VA visit date, provider type (physicians and physician assistants/nurse practitioners), patient demographics (age, gender, and race), clinical characteristics (Elixhauser comorbidity score), and fiscal month (October 2017 to September 2018).

Statistical Analyses

Descriptive statistics were calculated to explore the distributions of duplicate image ordering, the provider's rate of ordering images in the prior 6 months, the patient's Elixhauser comorbidity score, and other covariates. In our primary statistical model, we used a logistic regression to estimate the relationship between the provider's JLV use in the prior 6 months (yes/no) and duplicate imaging (yes/no). In an alternative specification, we used instrumental variables to focus on JLV use in the actual primary care visit. To deal with the potential endogeneity problem that unobserved patient characteristics might confound that relationship, we used a 2-stage residual inclusion (2SRI) logistic regression model to estimate the relationship between the provider's JLV use during the primary care visit (yes/no) and the ordering of duplicate imaging studies (yes/no) with JLV provider as the instrumental variable.

More formally, we wished to estimate the relationship between duplicate imaging and use of JLV during the primary care visit, controlling for potential confounders, including provider imaging rate in the prior 6 months, provider type (physician or physician assistant/nurse practitioner), patient age, gender, Elixhauser comorbidity risk score, time (month), and facility (VA Medical Center) (equation 1). This model could produce biased estimates because the decision to use JLV during the visit could be simultaneously determined with the decision to order a duplicate imaging study due to unobserved confounding factors such as patient complexity. To address this problem, we estimated the first stage model, which related JLV use during the visit to the provider's JLV use history (the JLV provider variable). Then, we estimated the second stage model, which related duplicate imaging to JLV use during the visit. Anscombe residuals ($X_{\mu e}$) calculated from the first stage were included in the second stage model because 2SRI models with Anscombe residuals generate less biased estimates for rare outcomes compared to 2SRI models with other forms of residuals [16]. Bootstrapping was used to improve the estimation of standard errors.

$$\text{2SRI logistic regression model: } Y = f(X_e\beta_e + X_o\beta_o) + X_\mu\beta_\mu + \varepsilon \quad \textbf{(1)}$$

Y: Provider ordering duplicate imaging

X_e: JLV encounter, endogenous

X_o: Provider imaging rate in prior 6 months, provider type (physician or physician assistant/nurse practitioner), patient age, gender, Elixhauser comorbidity risk score, time (month), and facility (VA Medical Center)

X_μ: Unobserved confounding factor such as patient complexity

ε: Residual

$$\text{First stage: } X_e = W\alpha + X_\mu\beta_\mu \quad \textbf{(2)}$$

W: The instrument of JLV provider and observed exogenous variables

$$\text{Second stage: } Y = f(X_e\beta_e + X_o\beta_o + X_{\mu e}\beta_\mu) + \varepsilon \quad \textbf{(3)}$$

$X_{\mu e}$: Anscombe residual calculated from the first stage model estimates

We tested the robustness of the result by using different model specifications, including ordinary least squares models on the relationship between JLV provider and duplicate imaging and linear 2-stage least squares models on the relationship between JLV encounter and duplicate imaging using JLV provider as an instrumental variable. We used Stata 15.1 (StataCorp LLC) to conduct the statistical analysis.

Results

Overall, JLV use has increased since fiscal year 2015. Rapid growth of monthly JLV audits was observed in fiscal year 2018 (Figure 1). Table 1 shows that the duplicate imaging rate among non–JLV encounters was 7.8% (34/435) and the duplicate imaging rate among JLV encounters was 7.9% (36/457). However, a direct comparison of these rates may be a biased estimate of the effect of JLV use owing to the endogeneity problem discussed above. The duplicate imaging rates were 11.2% (34/305) and 6.1% (36/587) among the non–JLV provider and JLV provider groups, respectively. Unlike the first comparison, this one should not be biased by uncontrolled differences in patient characteristics. Of the 892 unique patient-provider encounters in our analytic sample, 588 (65.9%) were males and the average age was 34.3 years, 512 (57.4%) patients were White, 228 (25.6%) were Black, and 152 (17%) were of other races. On average, patients had an Elixhauser comorbidity score of 1.2. Among the providers of these encounters, 336 (37.7%) were physicians and 556 (62.3%) were physician assistants/nurse practitioners. Providers had an average image ordering rate of 17.5% in the prior 6 months. Average patient and provider characteristics were not significantly different between the JLV encounter and non–JLV encounter groups or between the JLV provider and non–JLV provider groups (Table 1).

Figure 1. Joint Longitudinal Viewer use growth. JLV: Joint Longitudinal Viewer.

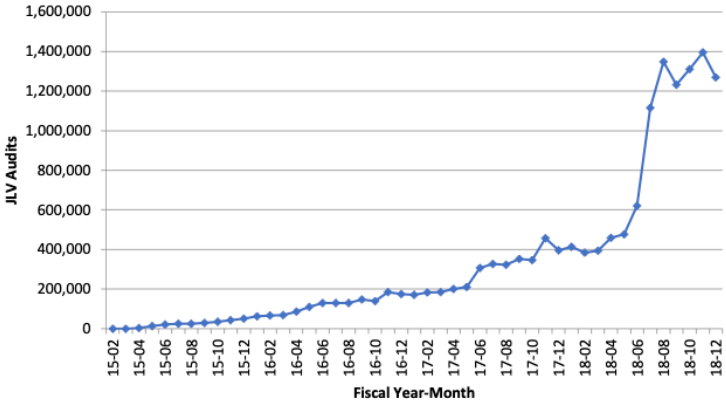


Table 1. Characteristics of the recently separated service members receiving Veterans Affairs primary care in fiscal year 2018 and characteristics of the related primary care providers (by Joint Longitudinal Viewer Encounter and Joint Longitudinal Viewer Provider).

Characteristics	Non–JLV[a] encounter (n=435)	JLV encounter (n=457)	Non–JLV provider (n=305)	JLV provider (n=587)	Overall (N=892)
Patient characteristics					
Gender (male), n (%)	298 (68.5)	290 (63.5)	208 (68.2)	380 (64.7)	588 (65.9)
Age (years), mean	34.2	34.3	34.5	34.1	34.3
Race, n (%)					
White	243 (55.9)	269 (58.9)	164 (53.8)	348 (59.3)	512 (57.4)
Black or African American	114 (26.2)	114 (24.9)	85 (27.9)	143 (24.4)	228 (25.6)
Other	78 (17.9)	74 (16.2)	56 (18.4)	96 (16.4)	152 (17)
Elixhauser comorbidity score (mean)	1.16	1.28	1.21	1.23	1.22
Provider characteristics					
Provider history rate of ordering imaging studies (mean)	17.8	17.2	18	17.3	17.5
Provider type, n (%)					
Physician	169 (38.9)	167 (36.5)	138 (45.3)	198 (33.7)	336 (37.7)
Physician assistant/Nurse practitioner	266 (61.1)	290 (63.5)	167 (54.8)	389 (66.3)	556 (62.3)

[a]JLV: Joint Longitudinal Viewer.

In our primary analysis with provider history of JLV utilization as the independent variable, after controlling for patient and provider characteristics, provider JLV use was significantly associated with a reduced likelihood (odds ratio [OR] 0.44, 95% CI 0.24-0.78; *P*=.005; average incremental effect=–0.05) of ordering duplicate image studies. Provider history of ordering images and patient Elixhauser comorbidity scores were strong confounders of the relationship between JLV use and duplicate imaging. Providers with high rates of ordering images in the prior 6 months were more likely to order duplicate images (OR 4.15, 95% CI 1.86-9.25; *P*=.001). Patient Elixhauser comorbidity scores of 3 or more were significantly associated with a reduced likelihood of receiving duplicate imaging services (OR 0.15, 95% CI 0.04-0.52; *P*=.003) (Table 2). In a logistic regression model, the average incremental effect is a nonlinear function of the coefficients and values of other explanatory variables [17]. Although the average incremental effect is easier to interpret than the OR, the statistical significance of the average incremental effect does not necessarily correspond to the significance of the coefficient or OR. As a result, the significance of the average incremental effect is not reported above.

In the 2SRI analysis, the results of our first stage model (Table 3) indicated that past use of JLV was strongly predictive of use of JLV by the provider during the primary care encounter (OR 1.43, 95% CI 1.05-1.81; *P*<.001). In a test of the coefficient on the instrument, a Cragg-Donald Wald F statistic greater than 10 indicates that the instrument is strong enough [17]. In a linear version of the first stage model, the strength of the instrument was tested (Cragg-Donald Wald F=43.68; *P*<.001). The Cragg-Donald Wald F statistic of 43.68 is greater than 10 and suggests that provider past use of JLV was a strong instrument. This is a necessary condition for JLV provider to serve as an instrumental variable for JLV encounter in our 2-stage specifications.

In the analysis assessing the relationship between JLV encounter and duplicate imaging with JLV provider as an instrumental variable, provider use of JLV was significantly associated with a reduction (OR 0.08, 95% CI 0.01-0.81; *P*=.03; average incremental effect=–0.16) in the likelihood of ordering duplicate images, controlling for patient and provider characteristics, time effects, and facility random effects (Table 4). Provider history of ordering images and patient Elixhauser comorbidity scores were strong confounders of the relationship between JLV use and duplicate imaging. Providers with high rates of ordering image studies in the prior 6 months were more likely to order duplicate images (OR 3.93, 95% CI 1.42-10.94; *P*=.009) compared to providers with low rates of ordering medical image studies. Patient Elixhauser comorbidity scores of 3 or more were significantly associated with a reduced likelihood of receiving duplicate imaging procedures (OR 0.16, 95% CI 0.04-0.57; *P*=.005) (Table 4).

Our main finding that JLV use had a significant effect on reducing duplicate imaging was robust using different model specifications, including 2-stage least squares models estimating the association between JLV encounter and duplicate imaging (see Table S2 in Multimedia Appendix 1).

Table 2. The impact of provider use of Joint Longitudinal Viewer in the prior 6 months of outpatient primary care visits on provider ordering of duplicate images.

Characteristics	Odds ratio (95% CI)[a]	*P* value
Provider characteristics		
Joint Longitudinal Viewer provider	0.44 (0.24-0.78)	.005
Provider history of ordering imaging studies (quartiles)		
1	Ref[b]	Ref
2	0.87 (0.33-2.32)	.78
3	2.73 (1.23-6.06)	.01
4	4.15 (1.86-9.25)	.001
Provider type		
Physician	Ref	Ref
Physician assistant/Nurse practitioner	1.32 (0.75-2.32)	.34
Patient characteristics		
Gender		
Female	Ref	Ref
Male	1.68 (0.89-3.17)	.11
Age (years)		
<30	Ref	Ref
30-39	2.19 (1.17-4.11)	.01
40-49	0.78 (0.35-1.73)	.54
≥50	1.27 (0.46-3.56)	.65
Race		
White	Ref	Ref
Black or African American	1.09 (0.56-2.13)	.80
Other	1.63 (0.82-3.21)	.16
Elixhauser comorbidity score		
0	Ref	Ref
1	0.44 (0.23-0.83)	.01
2	0.40 (0.18-0.90)	.03
3 and above	0.15 (0.04-0.52)	.003

[a]The odds ratio and 95% CIs are estimated from the logistic regression model controlling for all variables shown in the table as well as facility (random effects) and fiscal month.

[b]Ref indicates baseline in the analysis.

Table 3. The impact of provider use of Joint Longitudinal Viewer during outpatient primary care visits on provider ordering of duplicate images (stage 1 full output).[a]

Joint Longitudinal Viewer encounter (first stage)	Odds ratio (95% CI)	*P* value
Joint Longitudinal Viewer provider	1.43 (1.05 to 1.81)	<.001
Provider characteristics		
Provider history of ordering imaging studies (quartiles)		
1	Ref[b]	Ref
2	–0.18 (–0.63 to 0.28)	.45
3	0.30 (–0.17 to 0.78)	.21
4	–0.19 (–0.67 to 0.30)	.46
Provider type		
Physician	Ref	Ref
Physician assistant/Nurse practitioner	–0.06 (–0.42 to 0.30)	.75
Patient characteristics		
Gender		
Female	Ref	Ref
Male	–0.20 (–0.56 to 0.16)	.28
Age (years)		
<30	Ref	Ref
30-39	0.03 (–0.38 to 0.43)	.90
40-49	0.10 (–0.34 to 0.55)	.64
≥50	0.37 (–0.30 to 1.03)	.28
Race		
White	Ref	Ref
Black or African American	–0.30 (–0.70 to 0.11)	.15
Other	–0.29 (–0.75 to 0.18)	.22
Elixhauser comorbidity score		
0	Ref	Ref
1	–0.01 (–0.40 to 0.39)	.98
2	0.29 (–0.19 to 0.77)	.23
3 and above	0.13 (–0.37 to 0.64)	.61
Fiscal month		
1	Ref	Ref
2	0.15 (–0.59 to 0.90)	.69
3	–0.47 (–1.26 to 0.31)	.24
4	–0.63 (–1.42 to 0.17)	.12
5	–0.14 (–0.97 to 0.68)	.73
6	–0.49 (–1.27 to 0.29)	.22
7	0.62 (–0.15 to 1.39)	.12
8	0.61 (–0.10 to 1.33)	.09
9	0.65 (–0.07 to 1.37)	.08
10	0.59 (–0.16 to 1.35)	.12
11	0.68 (–0.10 to 1.46)	.09
12	0.77 (–0.09 to 1.63)	.08

Joint Longitudinal Viewer encounter (first stage)	Odds ratio (95% CI)	*P* value
Cons[c]	–1.11 (–1.88 to –0.33)	.005

[a]Average incremental effects are estimated from the 2-stage residual inclusion logistic regression controlling for all variables shown in the table.

[b]Ref indicates baseline in the analysis.

[c]Cons: Constant term in the regression.

Table 4. The impact of provider use of Joint Longitudinal Viewer during outpatient primary care visits on provider ordering of duplicate images (Stage 2 full output).[a]

Duplicate imaging (second stage)	Odds ratio (95% CI)	*P* value
Joint Longitudinal Viewer encounter	0.08 (0.01-0.81)	.03
Anscombe residual	3.16 (1.29-7.79)	.01
Provider characteristics		
Provider history of ordering imaging studies (quartiles)		
1	Ref[b]	Ref
2	0.83 (0.23-3.07)	.78
3	3.11 (1.18-8.22)	.02
4	3.93 (1.42-10.94)	.009
Provider type		
Physician	Ref	Ref
Physician assistant/Nurse practitioner	1.24 (0.61-2.51)	.56
Patient characteristics		
Gender		
Female	Ref	Ref
Male	1.49 (0.72-3.08)	.28
Age (years)		
<30	Ref	Ref
30-39	2.28 (0.98-5.34)	.06
40-49	0.87 (0.32-2.42)	.79
≥50	1.50 (0.38-5.93)	.57
Race		
White	Ref	Ref
Black or African American	1.03 (0.48-2.18)	.95
Other	1.57 (0.65-3.80)	.32
Elixhauser comorbidity score		
0	Ref	Ref
1	0.43 (0.20-0.91)	.03
2	0.43 (0.15-1.21)	.11
3 and above	0.16 (0.04-0.57)	.005
Fiscal month		
1	Ref	Ref
2	1.41 (0.00-602.30)	.91
3	1.31 (0.00-542.59)	.93
4	1.43 (0.00-639.45)	.91
5	0.75 (0.00-290.28)	.93
6	1.10 (0.00-523.39)	.98
7	3.15 (0.01-1539.80)	.72
8	1.77 (0.01-819.76)	.86
9	4.11 (0.01-1699.58)	.65
10	4.41 (0.01-2011.70)	.64
11	1.96 (0.01-762.17)	.83

Duplicate imaging (second stage)	Odds ratio (95% CI)	*P* value
12	5.02 (0.01-2177.41)	.60
Cons[c]	0.05 (0.00-22.75)	.34

[a]Average incremental effects are estimated from the 2-stage residual inclusion logistic regression controlling for all variables shown in the table.

[b]Ref indicates baseline in the analysis.

[c]Cons: Constant term in the regression.

Discussion

This study using national data from VA and DoD found that providers who viewed integrated patient health records from multiple settings were less likely to order potentially duplicate imaging studies for patients who had prior imaging studies conducted within 90 days. Based on results from our primary analysis, providers with a history of using JLV were 5 percentage points less likely to order duplicate images during a VA primary care visit for recently separated service members compared to providers who did not have a history of using JLV. Using the JLV provider as an independent variable, we were able to reduce potential endogeneity due to unobserved confounders that were associated with both JLV use during the primary care visit and ordering of duplicate images.

Our results were consistent with previous findings that use of HIE systems was associated with a reduction in repeat imaging studies [14] and that a longitudinal viewer of patient records from multiple sources was related to more positive patient experiences of care [7]. Our analysis had the added advantage of including provider-level variables, primarily a provider history of ordering images, which appeared to be a strong confounder. Access to national-level VA and DoD data also enabled the study to focus on images ordered for recently separated service members, who were transitioning between health care delivery systems and may be particularly likely to benefit from investments in integrated health information viewers or HIE systems.

Our study has several limitations. First, we focused on VA primary care visits and images within the 90-day follow-up period of a DoD image and therefore were unable to capture duplicate images in other settings such as community-based clinics. Further research could examine duplicate image studies ordered during different types of outpatient and inpatient encounters to improve the generalizability of the results. Second, limited by the administrative data source, we could not determine whether the identified duplicate imaging procedures were unnecessary. In some cases, providers may need to examine repeat image studies for serial changes in disease status or order follow-up imaging studies based on recommendations in the patient's previous imaging reports. Thus, some of the duplicate image studies we identified might have been clinically appropriate. We mitigated this limitation by excluding patients with cancer diagnoses and ensuring the consistency of the definition of duplicate imaging studies among the providers who used and did not use JLV. Third, restricted by data access, we could not adjust for HIE through another widely used health information viewer, VistaWeb, which was recently decommissioned. We tried to overcome this limitation by focusing on VA primary care visits in fiscal year 2018—the year when we observed rapid growth in JLV utilization after the VA's transition from VistaWeb to JLV. Fourth, our result was not robust when we changed the definition of JLV provider to providers with 10 or more JLV audits in the 6 months prior to the VA primary care visit, suggesting the heterogeneity of JLV benefits by frequency of use.

Organizational fragmentation and discontinuity of care have been linked to increased costs and adverse outcomes in VA and other health care settings [6]. Our findings suggest that the use of a longitudinal viewer of health records from multiple electronic health record sources has the potential to alleviate patient time burden, reduce adverse health effects of radiation, and decrease costs resulting from unnecessary duplicate imaging procedures. Health systems outside the VA could also consider investments in health information viewers or HIE technology to reduce the deleterious effects of fragmentation.

In conclusion, this study provides evidence that adoption of a longitudinal viewer of health records from multiple electronic health record systems is associated with a reduced likelihood of ordering duplicate image studies. In future studies, the association between health information viewers and other types of duplicate medical tests and care coordination metrics such as follow-up of suspicious lung nodules could be investigated to more fully illustrate the impact of HIE on quality and efficiency of care.

Acknowledgments

The authors would like to thank Linda Wedemeyer for helpful comments during this analysis. The Veterans Health Administration Office of Health Informatics provided financial support. Statements in this paper reflect the views of the authors and not necessarily the official positions of the US Department of Veterans Affairs or Boston University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full definitions of covariates and results from sensitivity analyses.
[DOCX File , 30 KB - medinform_v10i5e32168_app1.docx]

References

1. Vest JR, Gamm LD. Health information exchange: persistent challenges and new strategies. J Am Med Inform Assoc 2010;17(3):288-294 [FREE Full text] [doi: 10.1136/jamia.2010.003673] [Medline: 20442146]
2. Kuperman GJ. Health-information exchange: why are we doing it, and what are we doing? J Am Med Inform Assoc 2011;18(5):678-682 [FREE Full text] [doi: 10.1136/amiajnl-2010-000021] [Medline: 21676940]
3. Rudin RS, Motala A, Goldzweig CL, Shekelle PG. Usage and Effect of Health Information Exchange. Ann Intern Med 2014 Dec 02;161(11):803. [doi: 10.7326/m14-0877]
4. Kaelber DC, Bates DW. Health information exchange and patient safety. J Biomed Inform 2007 Dec;40(6 Suppl):S40-S45 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.011] [Medline: 17950041]
5. The Computer-Based Patient Record: An Essential Technology for Health Care. Washington, DC: National Academies Press; Dec 1997.
6. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005;Suppl Web Exclusives:W5-10. [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]
7. Legler A, Price M, Parikh M, Nebeker JR, Ward MC, Wedemeyer L, et al. Effect on VA Patient Satisfaction of Provider's Use of an Integrated Viewer of Multiple Electronic Health Records. J Gen Intern Med 2019 Jan;34(1):132-136 [FREE Full text] [doi: 10.1007/s11606-018-4708-z] [Medline: 30338474]
8. Chen M, Guo S, Tan X. Does Health Information Exchange Improve Patient Outcomes? Empirical Evidence From Florida Hospitals. Health Aff (Millwood) 2019 Feb;38(2):197-204. [doi: 10.1377/hlthaff.2018.05447] [Medline: 30715992]
9. Bailey JE, Pope RA, Elliott EC, Wan JY, Waters TM, Frisse ME. Health information exchange reduces repeated diagnostic imaging for back pain. Ann Emerg Med 2013 Jul;62(1):16-24. [doi: 10.1016/j.annemergmed.2013.01.006] [Medline: 23465552]
10. Pizer SD, Gardner JA. Is fragmented financing bad for your health? Inquiry 2011;48(2):109-122 [FREE Full text] [doi: 10.5034/inquiryjrnl_48.02.02] [Medline: 21898983]
11. Smith-Bindman R, Miglioretti DL, Johnson E, Lee C, Feigelson HS, Flynn M, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010. JAMA 2012 Jun 13;307(22):2400-2409 [FREE Full text] [doi: 10.1001/jama.2012.5960] [Medline: 22692172]
12. Lin EC. Radiation risk from medical imaging. Mayo Clin Proc 2010 Dec;85(12):1142-1146 [FREE Full text] [doi: 10.4065/mcp.2010.0260] [Medline: 21123642]
13. Einstein AJ. Medical imaging: the radiation issue. Nat Rev Cardiol 2009 Jun;6(6):436-438 [FREE Full text] [doi: 10.1038/nrcardio.2009.53] [Medline: 19471288]
14. Vest J, Kaushal R, Silver M, Hentel K, Kern L. Health information exchange and the frequency of repeat medical imaging. Am J Manag Care 2014 Nov;20(11 Spec No. 17):eSP16-eSP24 [FREE Full text] [Medline: 25811815]
15. VA and DOD need to address ongoing difficulties and better prepare for future integrations. United States Government Accountability Office. URL: https://www.gao.gov/assets/gao-16-280.pdf [accessed 2022-04-07]
16. Deb P, Norton E, Manning W. Health Econometrics Using Stata. College Station, TX: Stata Press; 2017.
17. Stock J, Yogo M. Testing for weak instruments in linear IV regression. In: Andrews DWK Identification and Inference for Econometric Models 2005:80-108. [doi: 10.1017/cbo9780511614491.006]

Abbreviations

2SRI: 2-stage residual inclusion
DoD: Department of Defense
HIE: health information exchange
JLV: Joint Longitudinal Viewer
OR: odds ratio
VA: Veterans Affairs

Edited by C Lovis; submitted 16.07.21; peer-reviewed by Y Tani, L Rusu; comments to author 04.02.22; revised version received 18.02.22; accepted 25.02.22; published 20.05.22.

Please cite as:
Yuan Y, Price M, Schmidt DF, Ward M, Nebeker J, Pizer S
Integrated Health Record Viewers and Reduction in Duplicate Medical Imaging: Retrospective Observational Analysis
JMIR Med Inform 2022;10(5):e32168
URL: https://medinform.jmir.org/2022/5/e32168
doi:10.2196/32168
PMID:35594070

Original Paper

The Architecture of a Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study

Julian Gruendner[1], MA, MSc; Noemi Deppenwiese[2], MSc; Michael Folz[3], Dipl Inf; Thomas Köhler[4]; Björn Kroll[5], PhD; Hans-Ulrich Prokosch[1], PhD; Lorenz Rosenau[5], MSc; Mathias Rühle[6], MSc; Marc-Anton Scheidl[1], MSc; Christina Schüttler[1], PhD; Brita Sedlmayr[7], PhD; Alexander Twrdik[6], MSc; Alexander Kiel[4,6*], BSc; Raphael W Majeed[8,9*], MSc

[1]Chair of Medical Informatics, Friedrich-Alexander University Erlangen-Nu¨rnberg, Erlangen, Germany
[2]Center of Medical Information and Communication Technology, University Hospital Erlangen, Erlangen, Germany
[3]Institute of Medical Informatics, Goethe University Frankfurt, Frankfurt am Main, Germany
[4]Federated Information Systems, German Cancer Research Center, Heidelberg, Germany
[5]IT Center for Clinical Research, University of Lübeck, Lübeck, Germany
[6]Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig, Germany
[7]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany
[8]Institute for Medical Informatics, University Clinic Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany
[9]Universities of Giessen and Marburg Lung Center, German Centre For Lung Research, Justus-Liebig University Giessen, Giessen, Germany
*these authors contributed equally

Corresponding Author:
Julian Gruendner, MA, MSc
Chair of Medical Informatics
Friedrich-Alexander University Erlangen-Nu¨rnberg
Wetterkreuz 15
Erlangen, 91058
Germany
Phone: 49 9131 8567787
Email: julian.gruendner@fau.de

Abstract

Background: An essential step in any medical research project after identifying the research question is to determine if there are sufficient patients available for a study and where to find them. Pursuing digital feasibility queries on available patient data registries has proven to be an excellent way of reusing existing real-world data sources. To support multicentric research, these feasibility queries should be designed and implemented to run across multiple sites and securely access local data. Working across hospitals usually involves working with different data formats and vocabularies. Recently, the Fast Healthcare Interoperability Resources (FHIR) standard was developed by Health Level Seven to address this concern and describe patient data in a standardized format. The Medical Informatics Initiative in Germany has committed to this standard and created data integration centers, which convert existing data into the FHIR format at each hospital. This partially solves the interoperability problem; however, a distributed feasibility query platform for the FHIR standard is still missing.

Objective: This study described the design and implementation of the components involved in creating a cross-hospital feasibility query platform for researchers based on FHIR resources. This effort was part of a large COVID-19 data exchange platform and was designed to be scalable for a broad range of patient data.

Methods: We analyzed and designed the abstract components necessary for a distributed feasibility query. This included a user interface for creating the query, backend with an ontology and terminology service, middleware for query distribution, and FHIR feasibility query execution service.

Results: We implemented the components described in the *Methods* section. The resulting solution was distributed to 33 German university hospitals. The functionality of the comprehensive network infrastructure was demonstrated using a test data set based on the German Corona Consensus Data Set. A performance test using specifically created synthetic data revealed the applicability

of our solution to data sets containing millions of FHIR resources. The solution can be easily deployed across hospitals and supports feasibility queries, combining multiple inclusion and exclusion criteria using standard Health Level Seven query languages such as Clinical Quality Language and FHIR Search. Developing a platform based on multiple microservices allowed us to create an extendable platform and support multiple Health Level Seven query languages and middleware components to allow integration with future directions of the Medical Informatics Initiative.

Conclusions: We designed and implemented a feasibility platform for distributed feasibility queries, which works directly on FHIR-formatted data and distributed it across 33 university hospitals in Germany. We showed that developing a feasibility platform directly on the FHIR standard is feasible.

(JMIR Med Inform 2022;10(5):e36709) doi:10.2196/36709

KEYWORDS

federated feasibility queries; FHIR; distributed analysis; feasibility study; HL7 FHIR; FHIR Search; CQL; COVID-19; pandemic; health data; query; patient data; consensus data set; medical informatics; Fast Healthcare Interoperability Resources

Introduction

Context

The COVID-19 pandemic has highlighted the critical need for all countries to strengthen their health data and information systems. Timely, credible, reliable, and actionable data ensure that political decisions are data-driven and facilitate understanding, monitoring, and forecasting [1]. Khan et al [2] have pointed out the need to strengthen national preparedness and the requirement that national public health institutes overcome practical challenges that affect timely access to and use of data. Their analysis identified that the availability of robust information systems that allow relevant data to be collected, shared, and analyzed sufficiently rapidly is needed to provide a timely local response to infectious disease outbreaks in the future [2].

In Germany, the nationally funded Medical Informatics Initiative (MII; funded by the Ministry of Education and Research—Bundesministerium für Bildung und Forschung) through 4 funded consortia (Data Integration for Future Medicine [DIFUTURE] [3], Heidelberg-Göttingen-Hanover Medical Informatics [HiGHmed] [4], Medical Informatics in Research and Care in University Medicine [MIRACUM] [5], and Smart Medical Information Technology for Healthcare [SMITH] [6]) has, in recent years, led to the establishment of data integration centers (DICs) in almost all 34 German university hospitals. These university hospitals created data sharing networks within their respective consortia. However, no overarching cross-consortia research data and feasibility portal existed as of spring 2020.

Need and Task

To tackle the COVID-19 challenges, the Bundesministerium für Bildung und Forschung has initiated the network of university medicine hospitals, which has launched 13 different projects, for example, to coordinate action plans and diagnostic and therapeutic strategies and to provide a comprehensive COVID-19 data exchange (CODEX) platform [7,8]. Decentralized data collection within the CODEX project was based on the German Corona Consensus Data Set (GECCO), a data set specifically designed to collect data on patients with COVID-19 for research [9].

To make real hospital GECCO data available, university hospitals used Fast Healthcare Interoperability Resources (FHIR) repositories within their MII DIC. To support feasibility studies as part of the German Portal for Medical Research Data (Deutsches Forschungsdatenportal für Gesundheit [FDPG]) and to identify the size of decentral available data sets based on dedicated cohort characterizations (eg, described by Doods et al [10], Soto-Rey et al [11], and Laaksonen et al [12]), we developed a central feasibility portal, securely connected to all German university hospital GECCO FHIR data repositories. For timely design and development, owing to the pandemic, it was imperative to build on tools and experiences from previous projects and align the design for later strategic integration of this feasibility portal into FDPG of the MII [13].

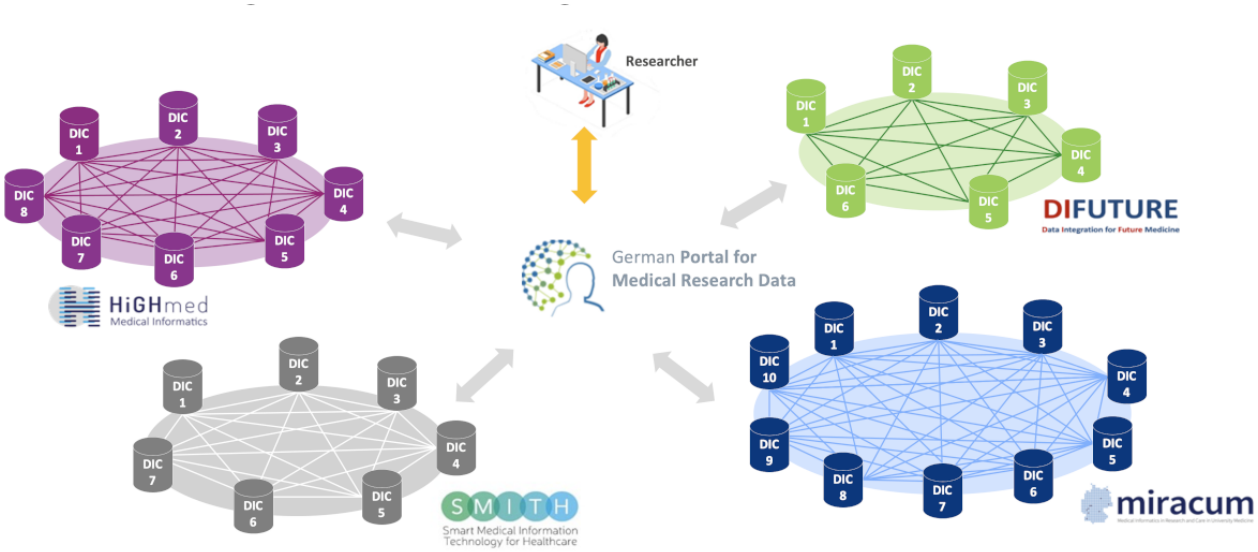
Background

First, the FDPG shall provide the central access point for researchers (Figure 1) to retrieve information about the availability of routine care data and biosamples in the network of all German university hospitals based on a central feasibility portal (which was, however, not yet developed in 2020). Second, it will provide functionality to electronically apply for data and biosample use in future projects. The latter functionality will manage all incoming research project applications, distribute these electronically to the DICs of all German university hospitals, and keep track of all application status replies from those decentral centers.

To allow studies to query and select patient data from a large, distributed pool of health care institutions, data need to be consolidated across these institutions. In contrast, the hospital landscape is very diverse, with each hospital using different systems and data formats. Although the 4 MII consortia have defined concepts for data harmonization within their consortia DIC (eg, openEHR in HiGHmed [14], the Informatics for Integrating Biology and the Bedside [i2b2] data model [5] in MIRACUM and DIFUTURE, Intersystems HealthShare in SMITH, and the Observational Medical Outcomes Partnership [OMOP] Common Data Model [15,16] in MIRACUM) within the MII, an agreement on a cross-consortia standardized data model was required. Thus, the emerging open standard FHIR [17], developed by Health Level Seven, is a promising candidate for addressing interoperability needs. Health care organizations widely adopt it to achieve interoperability, and it is increasingly supported by major electronic health record vendors. The rapidly

increasing availability of data in the FHIR format makes it a natural choice to collect real-world data, while allowing the possibility of translating it to other more specific formats in a relatively simple manner [18]. Thus, the MII working group for interoperability proposed the definition of the MII core data set [19] model based on FHIR. Consistent with this effort, the MII as a whole has agreed on FHIR as the de facto standard for interconsortia communication [20]. Therefore, every DIC in Germany has committed to make its data accessible via the FHIR standard application programming interface (API), making FHIR the only common format supported across all consortia.

Figure 1. Central German Portal for Medical Research Data and connection to all consortia and data integration centers (DICs; Medical Informatics Initiative).



Objectives

The objective of this study was to deduce and illustrate the conceptual design decisions for a distributed feasibility query portal directly based on FHIR data, including the underlying query transformation and execution tools and the middleware components implemented for secure network connections. We also aimed to describe the status of its implementation and use and provide an outlook on its future strategic integration in the German national MII infrastructure.

Methods

Abstract Architecture of a Distributed Feasibility Platform

A major challenge for the CODEX project was that any architecture should leverage the power of the German university hospital's DIC and be compatible with the agreed MII data sharing concepts. Thus, the CODEX project's feasibility portal was designed to serve as a generic basis for future developments in complementary MII projects. It was further conceived to be extendable to query the MII core data sets.

A feasibility query aims to identify suitable patients for a study. For feasibility, patient privacy can be guaranteed through anonymization by aggregation of the results, while still providing valuable information about the feasibility of a study, as only the number of patients is needed. The task of a distributed feasibility platform is to provide a user with the ability to specify a set of inclusion and exclusion criteria at a central location, send the query to participating sites, translate this query into a search query that can be executed inside a hospital's research data repository, and return the number of patients matching the criteria combination.

To achieve this, we had to create (1) a user interface (UI; feasibility UI) for creating and managing feasibility queries; (2) a backend service, which translates the user input into a standardized format (Structured Query) using an ontology service; (3) a middleware to securely transport the query; and (4) an execution service, which can process the standardized format, convert it to queries for an FHIR server, and execute the queries. Then, this service should return the number of patients identified.

Requirement Analysis and Architectural Design

The first step toward developing our tool was to define a list of capabilities (requirements) our platform should support. Building on previous studies on usability [21], query platforms [22], feasibility queries [23], and expert interviews, we curated and prioritized our requirements using Atlassian Confluence as collaboration platform [24]. The prioritization of the features was based on the added value of a feature and the potential estimated implementation cost. The identified features and their prioritization are presented in Multimedia Appendix 1.

The Structured Query as the central part of our feasibility process was developed across multiple meetings with the whole team, including experts on ontology, FHIR, FHIR Search, Clinical Quality Language (CQL), research data repositories, and medical data analysis. From the beginning, it was designed to provide a framework for feasibility queries, which, on the one hand, allowed to create feasibility queries across multiple grouped inclusion and exclusion criteria and, on the other hand, restricted the possible options in a way that makes it easy to translate it into existing FHIR query languages (CQL and FHIR

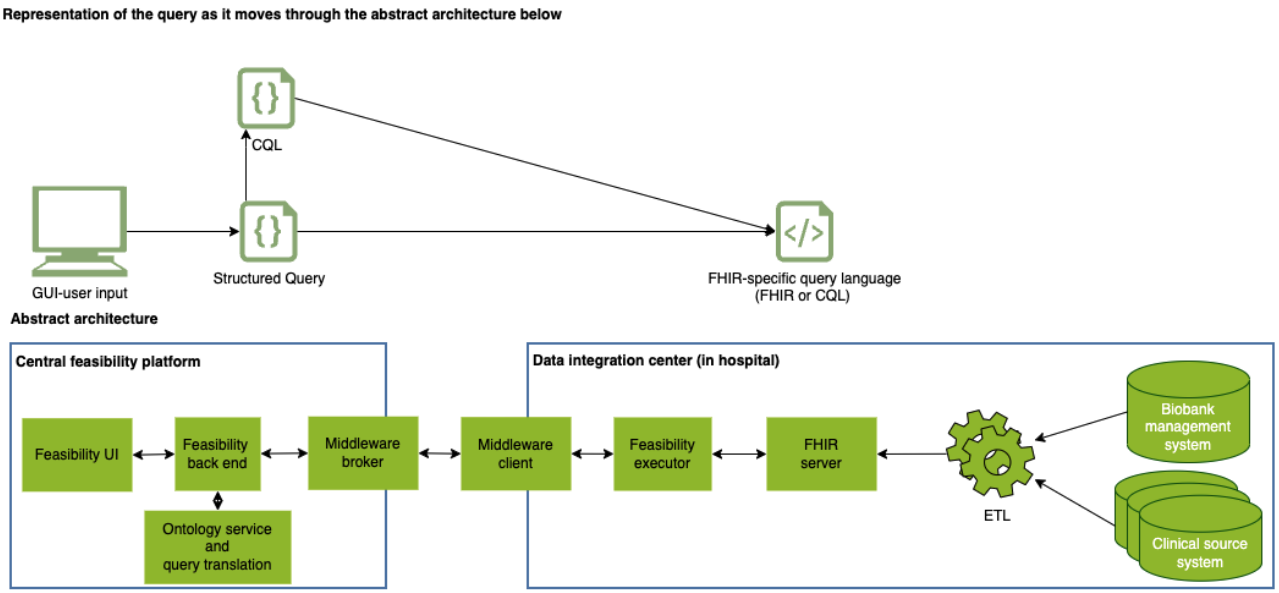
Search). The experts included expertise with existing query tools such as i2b2, OMOP, and Sample Locator [25], previously developed in other projects. This ensured that it would allow for capabilities similar to the existing query tools. The Structured Query, as evidenced by its specification [26], closely resembles the structure of the UI information, while providing sufficient abstraction to separate it from the UI by uniquely identifying single criteria based on their place within a given medical vocabulary. Building on the Structured Query and UI specifications, we worked closely with the whole team to define the necessary UI ontology (UI profiles) and a mapping file for query translation, which was to be used during query translation to enrich the basic definition of a criterion of the Structured Query with query language–specific parameters required for query translation. Critically, by analyzing the CQL language, we found that it has capabilities beyond the requirements of our feasibility specification, and therefore, we would have to specify a subset of CQL for query translation, leading to an incomplete translation, making it more fragile. Therefore, translation from a simpler (specifically restricted) format such as the Structured Query was considered to be easier and allowed us to control further development and separate the representation of the criteria from an implementation-specific system such as CQL and FHIR Search. Furthermore, the Structured Query, although independent of the UI, was designed to resemble it closely, making its generation by the UI easier, as the appropriate query object can be already built by the UI in JavaScript objects, which directly translate to the Structured Query in JSON format. Working across multiple institutions, we also had to consider how the queries and query results are securely exchanged between the different nodes of the network. This was achieved by using middleware components responsible for query transportation. To align the strategy with the other parts of the CODEX project and MII, we evaluated 4 middleware components as part of our project, which had been used previously to transport feasibility queries or used in other parts of the CODEX project to streamline further development. These included the AKTIN broker [27], data sharing framework (DSF) [28], connector component federated search [29], and German Biobank Node Client-Broker [30,31]. We then used the 2 middleware that had the highest scores as a base for further development. To calculate the score, 6 software developers from 5 institutions rated the existing solutions for code quality, documentation, complexity, and suitability for our requirements, on a scale of 1 (very good) to 5 (very bad).

Finally, based on experience from previous studies [22,32] and prototypes for data selection on FHIR servers, we knew that although CQL can support queries involving multiple criteria across different FHIR resources, the capability of FHIR Search is limited. Therefore, if FHIR Search was to be used for more complex queries, a software component was needed to execute and combine single FHIR Search queries to answer more sophisticated feasibility queries.

As part of our project, we performed a usability analysis based on our prototype implementation of the UI, the results of which were fed back into our development process to improve the user experience. This evaluation is described in more detail in a separate publication [33].

Figure 2 shows the abstract software components involved in the feasibility process. From left to right, it further illustrates how the representation of the query changes from user input, via a structured representation of the input (Structured Query) to an FHIR query language (FHIR Search [32,34,35] or CQL [16,36-40]) as it moves through the system.

Figure 2. Abstract software components of a distributed feasibility platform. CQL: Clinical Quality Language; ETL: extract-transform-load; FHIR: Fast Healthcare Interoperability Resources; GUI: graphical user interface; UI: user interface.



Performance Analysis

Performance of query execution depends on multiple factors including data set size, type of query execution (CQL vs FHIR Search), query composition (ie, number of criteria within a query), and number of resources processed as part of the query execution.

On the basis of these factors, we created 3 data sets (Table 1) with synthetic FHIR resources that would simulate different server loads and provide data sets, which would return a result for specific queries leading to query-specific data loads from 1, 10, 100, and 1000 thousands of patients. In addition, we augmented 2 of the data sets with background data, which consisted of 413,375 conditions (across 8593 unique condition codes), 270,505 procedures (across 6429 unique procedure codes), and 4,907,600 observations (across 1798 unique observation codes) to represent a typical distribution of data found across a hospital, based on the distributions of a German university hospital. This background data provide data within the server, which are not queried for, but might have an impact on index sizes and query execution speeds.

Furthermore, we created queries that included 4 criteria, each of which would be found exactly 1, 10, 100, or 1000 thousand times. The queries were designed to look for only 1 condition criterion (eg, ICD10–C50.1) or an AND combination of a patient, condition, procedure, and observation criterion (eg, female, ICD10 C50.1, OPS 5-787.ex, and LOINC 55782-7). The combination was always chosen to provide a specific load and has no clinical relevance. They were further chosen to demonstrate a worst-case scenario, where every part would have to be evaluated (AND rather than OR) to provide the answer, as every part would be true for this exact number of patients, implying that the program cannot terminate the search prematurely. We created CQL and Structured Queries for each query and, then, ran all CQL and Structured Queries on the same server 10 times consecutively after 1 warm-up run to ensure the same caching across each query. The host server had 8 cores, 16 GB RAM, and 320 GB solid state drive disk space. The repository for the performance test is available elsewhere [41].

Table 1. Performance test data sets.

Data set	Patients, n	Conditions, n	Procedures, n	Observations, n	Overall, n
Small	111,000	111,000	111,000	111,000	444,000
bg[a]-small	111,000	524,375	381,505	5,018,600	6,035,480
bg-large	1,111,000	1,524,375	1,381,505	6,018,600	10,035,480

[a]bg: background.

Results

Overview and Implementation

While implementing the abstract concept of a feasibility platform explained previously, reusing existing proven software artifacts from previous projects was a major requisite. The proposed architecture ensures strict modularity to achieve flexibility for future extensions and strategic alignments with other developments, for example, in MII. Finally, to fit into the existing architecture designs of the different MII consortia, partial duplication of modules and communication pathways (providing the university hospitals with optional implementation choices) for our development was accepted, when existing modular components could easily be integrated into a coherent framework. The detailed resulting architecture is illustrated in Figure 3.

The system's UI (feasibility UI) allows researchers to choose multiple criteria from an ontology tree (Figure 4) and combine them into a set of inclusion and exclusion criteria (Figure 5) using Boolean logic. The inclusion criteria are combined in a conjunctive normal form and the exclusion criteria in a disjunctive normal form.

Figure 3. Detailed architecture of the distributed feasibility platform. CQL: Clinical Quality Language; DSF: data sharing framework; ETL: extract-transform-load; FDPG: Deutsches Forschungsdatenportal für Gesundheit; FHIR: Fast Healthcare Interoperability Resources; FLARE: Feasibility Analysis Request Executor; UI: user interface.

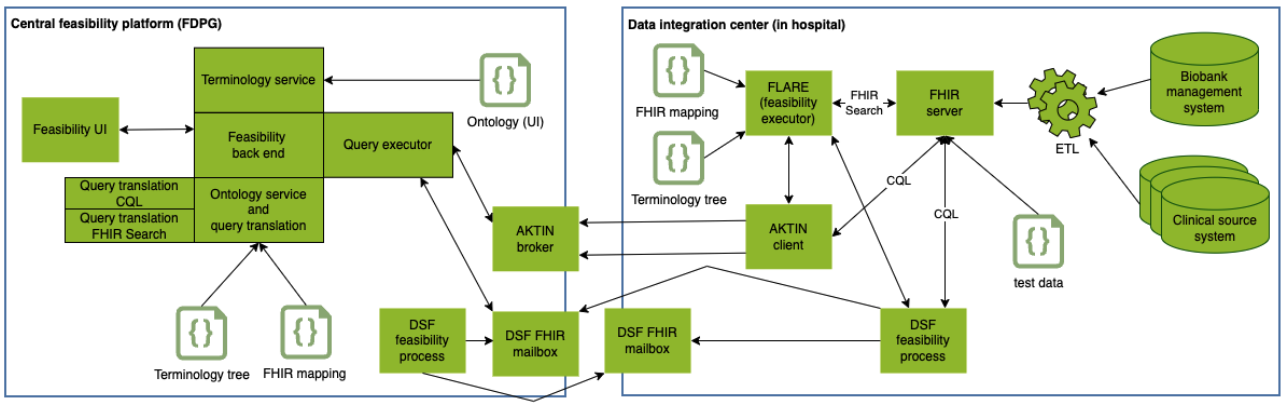
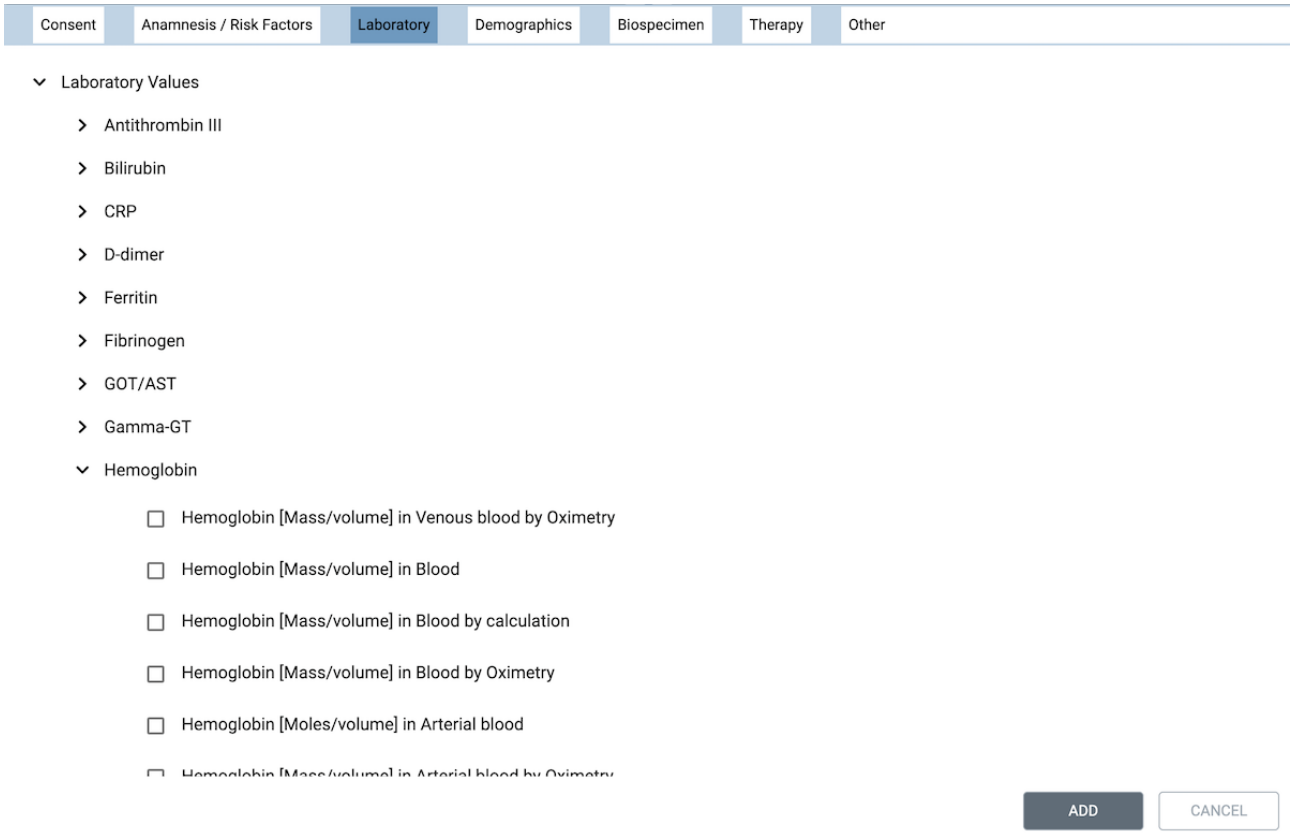


Figure 4. Example user interface representation of an ontology tree.

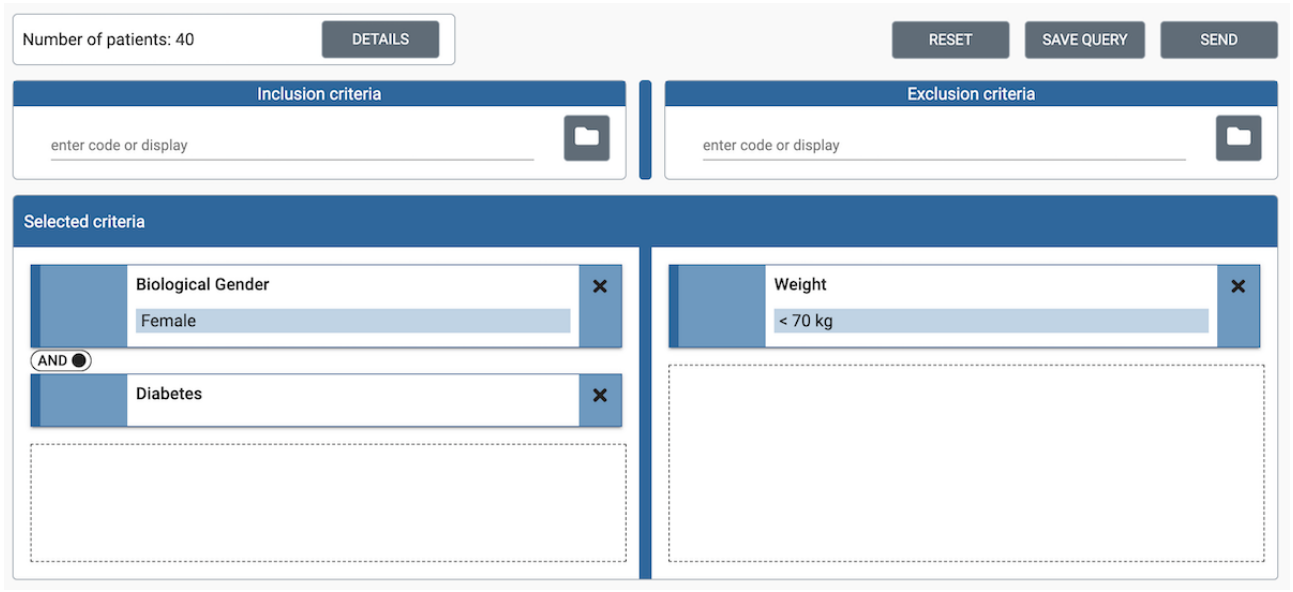


The ontology (ie, hierarchically structured concepts) for the UI is generated in JSON format based on the underlying FHIR profiles and a terminology service. A detailed description of how the ontology and mapping files are generated is described in a separate publication [42].

The process that generates the UI ontology also generates 2 configuration files (terminology tree and FHIR mapping). These files are required by the central feasibility backend and the decentral FHIR feasibility executor to process the input from the UI and translate it into FHIR-compatible search queries.

Once a researcher has created a feasibility query in the UI, it is converted into our Structured Query format. The Structured Query is a formal representation of the feasibility query, which structures the user input to allow easy translation into different query languages and closely resembles the user input structure. Currently, we support translation into 2 query languages used by FHIR servers: FHIR Search and CQL. Multimedia Appendix 2 illustrates the processing of the Structured Query example shown in Figure 5 from the UI to CQL and FHIR Search.

Figure 5. Example feasibility query in the user interface.



FHIR Search is part of the FHIR standard and implemented by most FHIR servers. However, currently, complex feasibility queries with intercriterion dependencies are not supported by FHIR Search. A way to overcome this limitation is to break a feasibility query into multiple smaller parts, each of which can be written as a single FHIR Search query. The parts (FHIR Search queries) are then sent to the FHIR server separately. The results are evaluated and combined using set algebra to calculate the final answer for a feasibility query.

For this purpose, we used the software library, Feasibility Analysis Request Executor (FLARE), initially developed for a research project by the University Hospital Rheinisch-Westfälische Technische Hochschule Aachen [43]. For our project, we contributed to the development of FLARE, by extending the software to support the Structured Query.

CQL is a high-level domain-specific query language, which is similar to Structured Query Language, built specifically with medical data in mind [39,40]. It supports, among many other use cases, the definition of cohort characterizations and counting of the respective cohort size, which are needed for feasibility queries. CQL is more powerful than FHIR Search; however, it is not as widely supported by current implementations of FHIR servers. Currently, the popularity of CQL is growing in the FHIR community and has recently been added to the HAPI project [37], a popular open-source FHIR reference implementation. Furthermore, it is supported by the Blaze FHIR server, developed within the German Biobank Alliance project [25], aimed at high-throughput performance. Blaze and the CQL language were chosen as an implementation option in the CODEX project after a comprehensive FHIR server benchmark. CQL has an advantage over FHIR Search in that even complex search queries can be written in a single query, which leads to faster query execution. Therefore, in CODEX, we support both CQL and FHIR Search.

To translate Structured Query to CQL and FHIR Search, we created translation components, which use an FHIR mapping JSON file to map each criterion to its respective FHIR query representation based on its coding (equivalent to FHIR coding type). The information provided by this mapping describes how the criterion is to be searched for inside the FHIR server. This includes the FHIR Search parameters to be used and the type of FHIR resource (eg, observation).

We use a terminology tree JSON file to find all the children of a criterion for inclusion in the respective search. This is necessary as researchers can select groups of criteria by selecting a parent criterion in a terminology hierarchy to include all child elements within the query. An example is the search for the diagnosis *Diabetes mellitus, Type 2*. If a researcher adds the diagnosis *Diabetes mellitus, Type 2* as a criterion (ICD10 code=E11) our tool expands the search to all subtypes of *Diabetes mellitus, Type 2* (including, for example, E11.0—*Diabetes mellitus, Type 2: with coma*). The information necessary to identify all subtypes of type 2 diabetes is provided in the terminology tree file.

A usability analysis of our prototype revealed that it is simple and intuitive. It also showed 26 problems, 8 of which were rated as “critical” [33]. However, usability problems were focused on the presentation of the UI or the ontology and will have no impact on the architectural decisions made. Specifically, our architecture will allow us to resolve these problems independent of the rest of the system for query translation, transportation, and execution.

Supporting Multiple Query Paths

CQL and FHIR Search have slightly different requirements regarding query execution. We generate multiple representations of the same feasibility query in the central feasibility backend and send all of them to each decentral DIC. This allows the DIC to configure which query representation to use without changing the central implementation. All feasibility queries can be generated as a single CQL query. Therefore, we generate the CQL query centrally and send this query to the DICs and their FHIR servers, which can execute them directly. As, in most cases, the FHIR Search representation of a Structured Query cannot be generated as a single query, we send the Structured Query to the DIC. Each DIC that prefers to use FHIR Search for the query execution will use the FLARE component locally. It translates each Structured Query into FHIR Search queries inside the respective DIC using the mapping and terminology tree files and executes them against the FHIR server.

Supporting Multiple Middleware

In our architecture, a middleware has the task to securely transport the query into a DIC and transport the answer to a query back to the central platform. In our case, the query is an object that contains the serialized version of our different query representations—Structured Query and CQL. To secure the connection between our central middleware components and local middleware clients, without requiring the university hospitals to open their firewall for outside requests, we chose a pull transport mechanism instead of a push process from the outside. Within CODEX, we evaluated multiple middleware components already developed by various MII partner sites and chose 2 that are already widely used in different consortia and fulfill the requirements mentioned previously: AKTIN broker [27,44] and HiGHmed DSF [28]. Both were extended to fully comply with the CODEX requirements, leading to a new client release for AKTIN [45] and the creation of a feasibility process for the DSF [46], similar to that created by Wettstein et al [47].

Privacy Through Anonymization by Aggregation and Access Restriction

The system we present here allows for querying patient data across multiple institutions from a central location, and information about patients is leaving the respective institution. This information, as any information about patients, is sensitive and needs to be anonymous when leaving an institution. The nature of feasibility queries is such that only an integer number leaves each participating hospital. However, a potential reconstruction of a patient profile by the central location would be possible if the exact number was returned. To avoid this, we aggregate each result by rounding it to the nearest 10 patients. A result of zero is returned as zero. We further restrict access to the platform to registered users and track all the created feasibility queries.

Containerization and Deployment Across Hospitals

The system described here is composed of many connected pieces of software, which must be installed across many institutions to create a feasibility query network of participating institutions. To ensure easy distribution of the software and streamline the installation process, we ensured that each software component created is distributed as a Docker image. We further tested our implementation for Kubernetes and installed a version of it in the Kubernetes cluster of the DIC of the university hospital in Erlangen, Germany. For easy installation across the institutions, we created an installation package, which provides an easy-to-install package based on multiple docker-compose files. In this first installation, the sites used only the AKTIN middleware, as the DSF was still in development and the set-up of the DSF proved to be more complex; for example, specific client certificates issued by an official certificate authority were required for its use. During the installation process, we found that the sites had very stringent firewalls and needed the option to support a proxy server between the client inside the hospital and the central broker. After adding proxy support, all the participating institutions could install the software and join the feasibility network.

First Ontology Generation and Test Across Hospitals

We implemented the architecture described previously and generated an ontology, a terminology tree, and a mapping file based on the GECCO FHIR profiles. We then distributed our implementation across the 33 participating institutions and asked them to load synthetic test data into their respective FHIR servers. We deployed the central feasibility tool and sent queries across the institutions. The test data set was generated based on synthetic data and converted to the MII FHIR format. We then used our UI to generate multiple test queries and found that we could create and execute them on our chosen FHIR servers. We further created a synthetic test patient data set in FHIR format [48] using the electronic data capture tool, REDCap (Research Electronic Data Capture; Vanderbilt University) [49], used by many participating institutions to capture COVID-19 data. The data set contains each type of criterion available in our UI. We verified our implementation using this data set.

Performance and Query Execution Speed

By running the performance tests (Table 2), we found that CQL was faster than FLARE as the number of resources processed increased. We also found that query execution time increased with the number of resources processed for a search and the amount of background data. For queries where small result sets had to be processed (<100,000 resources) and large amount of background data were loaded into the server, FLARE was faster than CQL. CQL processed all requests in <30 seconds. FLARE did not perform well with very large data sets and queries where >1,000,000 resources had to be processed, leading to execution times >47 seconds.

Table 2. Query response times across data set, query, and query execution type (CQL[a] and FLARE[b]).

Query	Criteria search for	Patients found, n	Resources processed, n	Response time by query execution and data set type (seconds), mean (SD) of 10 consecutive runs					
				cql-small	flare-small	cql-bg[c]-small	flare-bg-small	cql-bg-large	flare-bg-large
0	4	0	0	0.22 (0.01)	0.03 (0.0)	0.3 (0.01)	0.04 (0.0)	1.56 (0.04)	0.04 (0.0)
1000-1	1	1000	1000	0.57 (0.09)	0.11 (0.0)	0.85 (0.19)	0.11 (0.01)	5.52 (0.54)	0.13 (0.01)
1000-all	4	1000	4000	0.25 (0.03)	0.23 (0.06)	0.35 (0.05)	0.24 (0.06)	1.82 (0.06)	0.37 (0.26)
10000-1	1	10,000	10,000	0.56 (0.08)	0.5 (0.01)	0.89 (0.08)	0.49 (0.01)	5.49 (0.21)	0.67 (0.07)
10000-all	4	10,000	40,000	0.35 (0.04)	0.99 (0.08)	0.68 (0.07)	1.0 (0.1)	2.1 (0.08)	1.94 (1.37)
100000-1	1	100,000	100,000	0.85 (0.11)	4.34 (0.07)	1.13 (0.09)	5.16 (0.21)	6.07 (0.26)	5.37 (1.18)
100000-all	4	100,000	400,000	1.48 (0.12)	8.25 (0.41)	2.65 (0.23)	9.65 (0.24)	4.16 (0.18)	10.8 (0.09)
1000000-1	1	1,000,000	1,000,000	N/A[d]	N/A	N/A	N/A	10.49 (1.26)	47.53 (1.35)
1000000-all	4	1,000,000	4,000,000	N/A	N/A	N/A	N/A	29.05 (2.38)	119.64 (4.51)

[a]CQL: Clinical Quality Language.

[b]FLARE: Feasibility Analysis Request Executor.

[c]bg: background.

[d]N/A: not applicable.

Discussion

Principal Findings

We presented the concept and implementation of a distributed feasibility query platform, which works directly with FHIR-formatted hospital data. This demonstrates that the FHIR standard is suitable to build a feasibility platform on. FHIR Search does not support feasibility queries across multiple criteria directly. However, we built an FHIR feasibility executor, which combines single queries to answer these feasibility queries. This executor needs to load and combine the results of the different subqueries and, therefore, will be a performance bottleneck if single queries return large data sets. Therefore, we

also offer the fast but less widely available CQL query option. Furthermore, separating the concerns and supporting multiple query languages for query executions allows us to adjust to individual institutions' needs. Similarly, we found it to be useful to support multiple middleware components by providing clear interfaces. This supports more organizations and strategic directions and allows focusing on one middleware (AKTIN), while the other (DSF) is still being developed and deployed. Comparing the 2 middleware, the AKTIN implementation has the advantage of being simple, which is easy to maintain and extend for the purpose of transporting feasibility queries. It is agnostic to the query transported, so that the process extension necessary for the AKTIN implementation was easy and fast to achieve. Furthermore, the AKTIN middleware has been used successfully for several years in other projects. The DSF is an FHIR-based middleware and focuses on providing a platform for defining processes, which can be run across institutions. This enforces more structure than the AKTIN middleware, in the hope that this leads to improved interoperability. The DSF allows peer-to-peer communication if required. However, peer-to-peer communication is not relevant for feasibility queries from a central location. The biggest disadvantage of the DSF is that, with its large feature set, structure, and interoperability, it also introduces a high complexity to the system. Furthermore, the DSF is still in development and is yet to be used in a production environment. We chose to support both middleware in this project, as both have advantages and disadvantages, and the use of either within our future architecture largely depends on their respective use and acceptance within the MII.

Centering around the newly defined Structured Query format, which formally describes a feasibility query, allows the separation of the UI ontology from the translation into FHIR-compatible query languages. Therefore, the platform is built in a modular fashion and highly extendable. For example, one could imagine that an entirely different UI could be developed and integrated into the platform to satisfy future requirements, as long as it creates a Structured Query. Similarly, it allows the ontology, mapping, and what query execution languages the Structured Query is translated into, to be changed, to work with future query languages (eg, if the scope of the underlying data set changes). This allows the ontology for the front end to be created completely independent of the mapping and does not require a specific format for an ontology, allowing for quicker ontology generation compared with approaches that extend existing research platforms such as i2b2 [35]. The Structured Query can be considered as a new internal format for feasibility queries, and it could be argued that the representation as a Structured Query is not as interoperable as an FHIR representation. However, given the need to translate the query into multiple languages before being sent across institutions and that the Structured Query closely resembles the user input, the conversion from user input to Structured Query is much simpler than generating an analogous FHIR representation, which would then be converted again to FHIR Search and CQL. Furthermore, currently, no FHIR specification for feasibility queries exists, which would match the complexity of our Structured Query [32].

In the proposed architecture, the ontology and mapping to FHIR are added using the generated files. Thus, the used ontology and mapping to FHIR can be easily changed. This allows the feasibility platform to extend beyond our project and national boundaries. It is important to consider that any ontology used must be agreed by the institutions participating in a data sharing network and either be applicable directly or mapped at the decentral location according to the rules set by the institution. The FHIR standards' wide applicability, its wealth of complexity, and medical data entities it can support makes this a feasibility tool that can work with very diverse data, from laboratory data to conditions or biological specimen data. The translation and mapping we created is not restricted to a few FHIR resources, and the platform allows for the extension of the ontology and mapping to any FHIR resource. The fact that we generate mapping files, which can be distributed with our software, meant that the participating sites do not have to open an extra connection to a central terminology server or provide a terminology server themselves. This increases security and ease of installation.

Related Work

The FHIR standard has become more popular in recent years. More recently, it has been investigated not only for the exchange of patient data but also as a tool for data selection, extraction, and analysis [22,35]. With the popularity of the standard and the MII deciding to use FHIR as its main format for data exchange [19], the task was to build tools directly on the FHIR standard, rather than transforming data further to be analyzed with other software such as OMOP and i2b2 and tools built on their data models, such as Shared Health Research Information Network [50]. In this study, we designed and implemented a feasibility tool, which clearly separates the concerns of the different components and defines clear interfaces. This makes it easy to extend the platform and exchange components at each step of the process from user input to query execution and data storage. Similar to Paris et al [35], we present a feasibility platform, which works directly with the FHIR standard. Unlike Paris et al [35] we present a distributed system, which not only supports the translation of a query to FHIR Search but also the more powerful CQL query language. Hereby, we pave the way for translating standardized feasibility queries into other query languages based on structured input query, mapping, and term-code tree to resolve ontology hierarchies. Our implementation has the distinct advantage of allowing us to map user input to FHIR directly, rather than mapping user input to i2b2 objects and, then, to FHIR, thus reducing the overall complexity. Finally, as the usability of the existing feasibility UIs of i2b2 and OMOP can still be improved [21], the current architecture included and implemented a new and modern UI, which was found in our usability analysis to be intuitive and easy to use. Furthermore, the Sample Locator [25], previously developed as part of the German Biobank Alliance (originating from previous work in the German Cancer Consortium [51]) had the following limitations. (1) It did not include a generic terminology-based ontology tree, which allows researchers to select concepts easily. The current selection criteria were hard-coded, thus hindering the flexible extension of the UI. This is especially important, as the scope of the project will grow

over time. (2) It did not allow for more complex queries, such as grouping different criteria in OR groups within AND groups. Therefore, conjunctive (inclusion criteria) and disjunctive (exclusion criteria) normal form was chosen for the new UI, which supports more complex queries. (3) It allowed for direct *not* exclusion of criteria, which is currently not supported for all FHIR Search queries and would have introduced more complexity into the query translation process and had implications for performance (an FHIR Search for*: not* condition C50.1 would have returned a set of all other conditions). (4) It did not support time restrictions across all criteria.

The system's modular design supports different software components and final architecture decisions within the various MII university hospitals, depending on their local architecture design already existing within their DICs. Modularity with clearly defined APIs means that the comprehensive architecture framework can be adjusted easily, with locally preferred microservice components, if they fulfill the same functionality, thus supporting varying local requirements.

Beyond the analysis of the systems, as part of this study, there are many competing infrastructures for standardizing and distributing queries in a privacy-preserving manner. Specifically, for distributed analysis, multiple frameworks such as the Personal Health Train (PHT) [52,53] and DataSHIELD [54,55] exist. However, here, the focus is on a distributed feasibility platform for standardized feasibility queries that preserves privacy by aggregation at each site. This makes infrastructures such as the PHT and DataSHIELD, which focus on interactive and custom analyses, less well suited for our purpose. PHT specifically focuses on distributing custom analyses (algorithm+query) using containers to move the algorithms to the data. This is a great strength of the PHT, but it is not applicable for a structured feasibility query, which can be executed in the exact same manner (by the same algorithm) every time. Currently, the feasibility platform does not provide a mechanism for multiparty computing, allowing for exact responses to privacy-preserving feasibility queries across sites. This might be potentially relevant for rare diseases, where low numbers of patients would otherwise be returned to each site, thus making more accurate numbers essential. Previous work such as the PHT or DSF could be extended to provide a multiparty computing approach to return exact feasibility answers aggregated across multiple institutions. In the system described here, only the middleware would have to be replaced or extended, as the UI, query generation, and query execution at the sites would be identical.

Limitations

A feasibility platform across institutions works only if the institutions agree on the same ontology and map their data to the same terminologies or provide a mapping from the given input to their terminologies. In our project, we built on the German DIC data harmonization efforts. This ensures the compatibility of our queries with the data in each participating institution, as all DICs convert the data according to the same FHIR profiles and implementation guides of the MII core data set [56] and the GECCO [9] data set. Not all countries have these DICs, which means that extra data harmonization efforts would be required, which can be expensive and time-consuming. Furthermore, many electronic health record providers now support FHIR, but this does not necessarily mean that they provide the consented profiles or terminologies necessary for a distributed query. Creating a good ontology that is easy to use and provides the researcher with the right criteria is a difficult task. Many institutions generate ontologies manually, which means that they are carefully curated, but this is expensive and time-consuming. We successfully generated an ontology and mapping in an automatic process based on FHIR profiles and an ontology server. Whether this is applicable to arbitrary FHIR profiles still needs to be investigated.

The way we implemented the FHIR Search query path for multicriteria grouped feasibility queries means that the result sets of the sub-FHIR Search queries must be downloaded, patient IDs must be extracted, and the resulting sets must be combined. This download process may not be feasible for queries where parts return many results. To address this problem, currently, we also support CQL, which is a better option for large data sets. Our performance test demonstrated that CQL answers queries processing multiple millions of resources within 30 seconds. FLARE answered queries where 400,000 resources had to be processed in <12 seconds. Specifically, for COVID-19 data sets, we currently do not expect 1 site to return millions of patients, which means that the current implementation will answer queries on patients who are specific to COVID-19 in seconds rather than minutes. Furthermore, the finding that the number of resources processed is the main predictor of query execution time paves the way for future improvements. The current performance test, as well as being repeatable, allows one to draw conclusions on feasible data set sizes. However, a more comprehensive investigation with data sets of 200 or 500 million resources and different server sizes and better understanding of what large real-world data sets look like are still missing. This is especially relevant within the MII if the current feasibility portal is extended beyond the COVID-19 data set to analyze multiple years of real-world hospital data.

Future Directions and Conclusions

We presented the design and implementation of a feasibility platform for distributed feasibility queries, which works directly on FHIR-formatted data. The platform was deployed across 33 university hospitals and the viability of the approach was demonstrated using a set of synthetic test data in the appropriate format. Supporting FHIR Search directly requires a feasibility executor (FLARE) to answer feasibility queries across multiple criteria. The advantage of the FLARE approach is that it did not only overcome current FHIR Search limitations but will also provide a solution to further limitations in the future. An example of this is the implementation of time-dependent intercriterion relationships (eg, a specific laboratory value within 3 days of a medication), which we plan to implement in the future. This is possible as full FHIR resources can be processed, including the appropriate time stamp field for each resource, which can then be compared for the specified interresource time constraints for each patient. Our performance analysis revealed that our implemented feasibility platform can answer queries for large data sets (multiple millions of resources) within seconds and that CQL is significantly faster than FLARE. The

performance depends heavily on the number of patients for CQL and, for FLARE, the number of hits for each single criterion searched for. Consistent with this, we are planning to improve the performance of our implementation by using heuristics on the FHIR server to optimize FLARE and CQL query execution. This means identifying the criterion with the statistically lowest number of occurrences first, and then, querying further criteria with the reduced patient set. This is possible for FLARE and CQL and will be investigated by our team in the future. The implementation and design described here focused on the GECCO COVID-19 data set. The platform presented here is very generalizable and can be applied to any FHIR-formatted data or even to different query languages currently supported by different FHIR servers. One of the next steps is to integrate more data from different sources. In this pursuit, partners from the MII and German Biobank Alliance [57] have joined forces in 2021 to bring together previously independent initiatives for data and biosample sharing, by aligning information technology infrastructures and the respective regulatory and governance frameworks established in Germany within the biobanking community on one side and the medical informatics community on the other. The resulting Aligning Biobanks and DIC Efficiently [58] project started in May 2021 [59]. In our implementation, we only tested specific FHIR servers; however, our support of FHIR Search allows us to work with any standard FHIR API. Thus, testing our system with FHIR-APIs built on optimized database systems, as suggested by Paris et al [35], would be of interest. The platform presented here provides a solution only for the first part of the research cycle. Given the way the platform is built, currently, we return only the number of patients. One can easily imagine changing the return value to a list of patient IDs, which would allow the platform to create a cohort or patient subpopulation for a later decentral data selection process. This decentral cohort-creation process can then be combined with a decentral data selection process. This would allow a researcher to create a feature (criteria) set of data, based on the previously created cohort, which the researcher would like to extract for further analysis. Such a tool can then extract the required data and create a prepared data set for analysis at each site. In the simplest case, this prepared data set can be a comma-separated list of selected features for each patient. Creating such a tool would allow the FHIR standard to support distributed privacy-preserving analysis using tools such as DataSHIELD [54,60].

Acknowledgments

The authors thank all COVID-19 data exchange partners from the university hospitals in the network university medicine. This study 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 (grant 01KX2021).

Source Code Availability

The source code of the project is available on GitHub [61].

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of identified requirements for a feasibility platform.
[PDF File (Adobe PDF File), 118 KB - medinform_v10i5e36709_app1.pdf]

Multimedia Appendix 2
Representation of a feasibility query as it moves through the system from user interface representation, via Structured Query to Clinical Quality Language and Fast Healthcare Interoperability Resources Search.
[PDF File (Adobe PDF File), 177 KB - medinform_v10i5e36709_app2.pdf]

References

1. Azzopardi-Muscat N, Kluge HH, Asma S, Novillo-Ortiz D. A call to strengthen data in response to COVID-19 and beyond. J Am Med Inform Assoc 2021 Mar 01;28(3):638-639 [FREE Full text] [doi: 10.1093/jamia/ocaa308] [Medline: 33275146]
2. Khan MS, Dar O, Erondu NA, Rahman-Shepherd A, Hollmann L, Ihekweazu C, et al. Using critical information to strengthen pandemic preparedness: the role of national public health agencies. BMJ Glob Health 2020 Sep;5(9):e002830 [FREE Full text] [doi: 10.1136/bmjgh-2020-002830] [Medline: 32994228]
3. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn KA. Data Integration for Future Medicine (DIFUTURE). Methods Inf Med 2018 Jul;57(S 01):e57-e65 [FREE Full text] [doi: 10.3414/ME17-02-0022] [Medline: 30016812]
4. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed - an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018 Jul;57(S 01):e66-e81 [FREE Full text] [doi: 10.3414/ME18-02-0002] [Medline: 30016813]

5. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med 2018 Jul;57(S 01):e82-e91 [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
6. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med 2018 Jul;57(S 01):e92-105 [FREE Full text] [doi: 10.3414/ME18-02-0004] [Medline: 30016815]
7. CODEX | COVID-19 Data Exchange Platform. Medical Informatics Initiative Germany. URL: https://www.medizininformatik-initiative.de/en/node/588 [accessed 2022-01-13]
8. Prokosch HU, Bahls T, Bialke M, Eils J, Fegeler C, Gruendner J, et al. The COVID-19 data exchange platform of the German university medicine. Stud Heal Technol informatics Heal Inf (forthcoming) 2022.
9. 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]
10. Doods J, Bache R, McGilchrist M, Daniel C, Dugas M, Fritz F, Work Package 7. Piloting the EHR4CR feasibility platform across Europe. Methods Inf Med 2014;53(4):264-268. [doi: 10.3414/ME13-01-0134] [Medline: 24954881]
11. Soto-Rey I, Trinczek B, Amo JI, Bauselas J, Dugas M, Fritz F. Web-based multi-site feasibility questionnaire tool. Stud Health Technol Inform 2015;212:88-93. [Medline: 26063262]
12. Laaksonen N, Varjonen JM, Blomster M, Palomäki A, Vasankari T, Airaksinen J, et al. Assessing an Electronic Health Record research platform for identification of clinical trial participants. Contemp Clin Trials Commun 2020 Dec 18;21:100692 [FREE Full text] [doi: 10.1016/j.conctc.2020.100692] [Medline: 33409423]
13. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018 Jul;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
14. Wulff A, Sommer KK, Ballout S, HiGHmed Consortium, Haarbrandt B, Gietzelt M. A report on archetype modelling in a Nationwide Data Infrastructure Project. Stud Health Technol Inform 2019;258:146-150. [Medline: 30942733]
15. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform 2018 Jan;9(1):54-61 [FREE Full text] [doi: 10.1055/s-0037-1617452] [Medline: 29365340]
16. Reinecke I, Gulden C, Kümmel M, Nassirian A, Blasini R, Sedlmayr M. Design for a modular clinical trial recruitment support system based on FHIR and OMOP. Stud Health Technol Inform 2020 Jun 16;270:158-162. [doi: 10.3233/SHTI200142] [Medline: 32570366]
17. HL7 FHIR. URL: https://www.hl7.org/fhir/ [accessed 2020-02-04]
18. González-Castro L, Cal-González VM, Del Fiol G, López-Nores M. CASIDE: a data model for interoperable cancer survivorship information based on FHIR. J Biomed Inform 2021 Dec;124:103953 [FREE Full text] [doi: 10.1016/j.jbi.2021.103953] [Medline: 34781009]
19. Kerndatensatzmodule Person, Medikation und Laborbefund der Medizininformatik-Initiative. Medizin Informatik Initiative. URL: https://www.medizininformatik-initiative.de/de/kommentierung-der-kerndatensatzmodule-person-medikation-und-laborbefund-der-medizininformatik [accessed 2022-01-13]
20. Medizininformatik-Initiative beschließt Verwendung von FHIR. Medizin Informatik Initiative. 2010 Jul 5. URL: https://www.medizininformatik-initiative.de/de/medizininformatik-initiative-beschliesst-verwendung-von-fhir [accessed 2022-03-05]
21. Schüttler C, Prokosch HU, Sedlmayr M, Sedlmayr B. Evaluation of three feasibility tools for identifying patient data and biospecimen availability: comparative usability study. JMIR Med Inform 2021 Jul 21;9(7):e25531 [FREE Full text] [doi: 10.2196/25531] [Medline: 34287211]
22. Gruendner J, Gulden C, Kampf M, Mate S, Prokosch HU, 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]
23. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit Transl Bioinform 2010 Mar 01;2010:46-50 [FREE Full text] [Medline: 21347148]
24. Confluence | Your Remote-Friendly Team Workspace. Atlassian. URL: https://www.atlassian.com/software/confluence [accessed 2022-03-15]
25. Schüttler C, Prokosch HU, 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]
26. codex-structured-query/structured-query/example-json/2021_10_18_StructuredQueryV2Example.json. GitHub. 2021 Dec 10. URL: https://github.com/num-codex/codex-structured-query/blob/structured-query-v2/structured-query/example-json/2021_10_18_StructuredQueryV2Example.json [accessed 2022-03-10]
27. AKTIN search broker components: asynchronous distribution of search queries across federated data warehouses. GitHub. URL: https://github.com/aktin/broker [accessed 2021-12-07]
28. Hund H, Wettstein R, Heidt CM, Fegeler C. Executing distributed healthcare and research processes - the HiGHmed data sharing framework. Stud Health Technol Inform 2021 May 24;278:126-133. [doi: 10.3233/SHTI210060] [Medline: 34042885]

29. 66. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 12. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF). German Medical Science. 2021 Sep 30. URL: https://www.egms.de/static/en/meetings/gmds2021/21gmds070.shtml [accessed 2022-03-23]
30. samply/share-client. GitHub. URL: https://github.com/samply/share-client [accessed 2022-03-23]
31. samply/searchbroker. GitHub. URL: https://github.com/samply/searchbroker [accessed 2022-03-23]
32. Gulden C, Mate S, Prokosch HU, Kraus S. Investigating the capabilities of FHIR search for clinical trial phenotyping. Stud Health Technol Inform 2018;253:3-7. [Medline: 30147028]
33. Sedlmayr B, Sedlmayr M, Kroll B, Prokosch HU, Gruendner J, Schüttler C. Improving COVID-19 research of university hospitals in Germany: formative usability evaluation of the CODEX feasibility portal. Appl Clin Inform 2022 Mar;13(2):400-409. [doi: 10.1055/s-0042-1744549] [Medline: 35445386]
34. FHIR search. HL7 FHIR. URL: https://www.hl7.org/fhir/search.html [accessed 2020-06-15]
35. Paris N, Mendis M, Daniel C, Murphy S, Tannier X, Zweigenbaum P. i2b2 implemented over SMART-on-FHIR. AMIA Jt Summits Transl Sci Proc 2018 May 18;2017:369-378 [FREE Full text] [Medline: 29888095]
36. Clinical Quality Language (CQL). Clinical Quality Language Release 1. URL: https://cql.hl7.org/ [accessed 2021-12-07]
37. 8.0CQL Getting Started. HAPI FHIR Documentation. URL: https://hapifhir.io/hapi-fhir/docs/server_jpa_cql/cql.html [accessed 2022-01-13]
38. Kiel A. samply/blaze. GitHub. URL: https://github.com/samply/blaze [accessed 2020-05-12]
39. Nguyen BP, Reese T, Decker S, Malone D, Boyce RD, Beyan O. Implementation of clinical decision support services to detect potential drug-drug interaction using clinical quality language. Stud Health Technol Inform 2019 Aug 21;264:724-728. [doi: 10.3233/SHTI190318] [Medline: 31438019]
40. Soares A, Jenders RA, Harrison R, Schilling LM. A comparison of Arden syntax and clinical quality language as knowledge representation formalisms for clinical decision support. Appl Clin Inform 2021 May;12(3):495-506 [FREE Full text] [doi: 10.1055/s-0041-1731001] [Medline: 34192772]
41. num-codex/feasibility-performance-test. GitHub. URL: https://github.com/num-codex/feasibility-performance-test [accessed 2022-03-10]
42. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, et al. 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 Apr 27;10(4):e35789 [FREE Full text] [doi: 10.2196/35789] [Medline: 35380548]
43. rwth-imi/flare-query. GitHub. URL: https://github.com/rwth-imi/flare-query [accessed 2022-01-13]
44. Brammen D, Greiner F, Kulla M, Otto R, Schirrmeister W, Thun S, AKTIN-Notaufnahmeregister. [AKTIN - The German Emergency Department Data Registry - real-time data from emergency medicine : implementation and first results from 15 emergency departments with focus on Federal Joint Committee's guidelines on acuity assessment]. Med Klin Intensivmed Notfmed 2022 Feb;117(1):24-33 [FREE Full text] [doi: 10.1007/s00063-020-00764-2] [Medline: 33346852]
45. num-codex/codex-aktin-broker. GitHub. URL: https://github.com/num-codex/codex-aktin-broker [accessed 2021-12-07]
46. num-codex/codex-processes-ap2. GitHub. URL: https://github.com/num-codex/codex-processes-ap2 [accessed 2021-12-07]
47. Wettstein R, Hund H, Kobylinski I, Fegeler C, Heinze O. Feasibility queries in distributed architectures - concept and implementation in HiGHmed. Stud Health Technol Inform 2021 May 24;278:134-141. [doi: 10.3233/SHTI210061] [Medline: 34042886]
48. num-codex/codex-testdata-to-sq. GitHub. URL: https://github.com/num-codex/codex-testdata-to-sq/blob/main/testData.json [accessed 2021-12-07]
49. REDCap. URL: https://www.project-redcap.org/ [accessed 2021-12-07]
50. 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;16(5):624-630 [FREE Full text] [doi: 10.1197/jamia.M3191] [Medline: 19567788]
51. Joos S, Nettelbeck DM, Reil-Held A, Engelmann K, Moosmann A, Eggert A, et al. German Cancer Consortium (DKTK) - a national consortium for translational cancer research. Mol Oncol 2019 Mar;13(3):535-542 [FREE Full text] [doi: 10.1002/1878-0261.12430] [Medline: 30561127]
52. Choudhury A, van Soest J, Nayak S, Dekker A. Personal health train on FHIR: a privacy preserving federated approach for analyzing FAIR data in healthcare. In: Proceedings of the 2nd International Conference on Machine Learning, Image Processing, Network Security and Data Sciences. 2020 Presented at: MIND '20; July 30-31, 2020; Silchar, India p. 85-95. [doi: 10.1007/978-981-15-6315-7_7]
53. Beyan O, Choudhury A, van Soest J, Kohlbacher O, Zimmermann L, Stenzhorn H, et al. Distributed analytics on sensitive medical data: the personal health train. Data Intell 2020 Jan;2(1-2):96-107. [doi: 10.1162/dint_a_00032]
54. Gruendner J, Prokosch HU, Schindler S, Lenz S, Binder H. A queue-poll extension and DataSHIELD: standardised, monitored, indirect and secure access to sensitive data. Stud Health Technol Inform 2019;258:115-119. [Medline: 30942726]
55. Marcon Y, Bishop T, Avraam D, Escriba-Montagut X, Ryser-Welch P, Wheater S, et al. Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. PLoS Comput Biol 2021 Mar 30;17(3):e1008880 [FREE Full text] [doi: 10.1371/journal.pcbi.1008880] [Medline: 33784300]

56. Medizininformatik Initiative - KDS - Meta. SIMPLIFIER. URL: https://simplifier.net/MedizininformatikInitiative-Kerndatensatz/~introduction [accessed 2022-01-13]
57. Klingler C, von Jagwitz-Biegnitz M, Hartung ML, Hummel M, Specht C. Evaluating the German biobank node as coordinating institution of the German biobank alliance: engaging with stakeholders via survey research. Biopreserv Biobank 2020 Apr;18(2):64-72. [doi: 10.1089/bio.2019.0060] [Medline: 31859533]
58. Prokosch H, Baber R, Bollmann P, Gebhardt M, Gruendner J, Hummel M. Aligning biobanks and data integration centers efficiently (ABIDE_MI). Stud Health Technol Inform 2022 May 16;292:37-42. [doi: 10.3233/SHTI220317] [Medline: 35575846]
59. ABIDE_MI. Medizin Informatik Initiative. URL: https://www.medizininformatik-initiative.de/de/use-cases-und-projekte/abidemi [accessed 2021-12-03]
60. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol 2014 Dec;43(6):1929-1944 [FREE Full text] [doi: 10.1093/ije/dyu188] [Medline: 25261970]
61. num-codex/codex-develop. GitHub. 2021 Oct 22. URL: https://github.com/num-codex/codex-develop [accessed 2021-12-07]

Abbreviations

API: application programming interface
CODEX: COVID-19 data exchange
CQL: Clinical Quality Language
DIC: data integration center
DSF: data sharing framework
FDPG: Deutsches Forschungsdatenportal für Gesundheit
FHIR: Fast Healthcare Interoperability Resources
FLARE: Feasibility Analysis Request Executor
GECCO: German Corona Consensus Data Set
HiGHmed: Heidelberg-Göttingen-Hanover Medical Informatics
i2b2: Informatics for Integrating Biology and the Bedside
MII: Medical Informatics Initiative
MIRACUM: Medical Informatics in Research and Care in University Medicine
OMOP: Observational Medical Outcomes Partnership
PHT: Personal Health Train
REDCap: Research Electronic Data Capture
UI: user interface

Edited by C Lovis; submitted 24.01.22; peer-reviewed by P Holub, R Saripalle, S Meister; comments to author 23.02.22; revised version received 16.03.22; accepted 11.04.22; published 25.05.22.

Please cite as:
Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch HU, Rosenau L, Rühle M, Scheidl MA, Schüttler C, Sedlmayr B, Twrdik A, Kiel A, Majeed RW
The Architecture of a Feasibility Query Portal for Distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) Patient Data Repositories: Design and Implementation Study
JMIR Med Inform 2022;10(5):e36709
URL: https://medinform.jmir.org/2022/5/e36709
doi:10.2196/36709
PMID:35486893

Original Paper

User Perceptions and Use of an Enhanced Electronic Health Record in Rwanda With and Without Clinical Alerts: Cross-sectional Survey

Hamish S F Fraser[1], MBChB, MSc, FACMI; Michael Mugisha[2], BSc, MPH; Eric Remera[3], BSc, MSc; Joseph Lune Ngenzi[2], BSc, MSc; Janise Richards[4], MSc, MPH, PhD; Xenophon Santas[4], BA; Wayne Naidoo[5], BSc, MSc; Christopher Seebregts[5], BSc, PhD; Jeanine Condo[2], MSc, MD, PhD; Aline Umubyeyi[2], MSc, MD, PhD

[1]Brown Center for Biomedical Informatics, Brown University, Providence, RI, United States

[2]School of Public Health, University of Rwanda, Kigali, Rwanda

[3]Ministry of Health, Kigali, Rwanda

[4]Centers for Disease Control, Atlanta, GA, United States

[5]Jembi Health Systems, Cape Town, South Africa

Corresponding Author:
Hamish S F Fraser, MBChB, MSc, FACMI
Brown Center for Biomedical Informatics
Brown University
233 Richmond street
Providence, RI, 02912
United States
Phone: 1 401 863 1815
Email: hamish_fraser@brown.edu

Abstract

Background: Electronic health records (EHRs) have been implemented in many low-resource settings but lack strong evidence for usability, use, user confidence, scalability, and sustainability.

Objective: This study aimed to evaluate staff use and perceptions of an EHR widely used for HIV care in >300 health facilities in Rwanda, providing evidence on factors influencing current performance, scalability, and sustainability.

Methods: A randomized, cross-sectional, structured interview survey of health center staff was designed to assess functionality, use, and attitudes toward the EHR and clinical alerts. This study used the associated randomized clinical trial study sample (56/112, 50% sites received an enhanced EHR), pulling 27 (50%) sites from each group. Free-text comments were analyzed thematically using inductive coding.

Results: Of the 100 participants, 90 (90% response rate) were interviewed at 54 health centers: 44 (49%) participants were clinical and 46 (51%) were technical. The EHR top uses were to access client data easily or quickly (62/90, 69%), update patient records (56/89, 63%), create new patient records (49/88, 56%), generate various reports (38/85, 45%), and review previous records (43/89, 48%). In addition, >90% (81/90) of respondents agreed that the EHR made it easier to make informed decisions, was worth using, and has improved patient information quality. Regarding availability, (66/88) 75% said they could *always or almost always* count on the EHR being available, whereas (6/88) 7% said *never/almost never*. In intervention sites, staff were significantly more likely to update existing records (P=.04), generate summaries before (P<.001) or during visits (P=.01), and agree that “the EHR provides useful alerts, and reminders” (P<.01).

Conclusions: Most users perceived the EHR as well accepted, appropriate, and effective for use in low-resource settings despite infrastructure limitation in 25% (22/88) of the sites. The implementation of EHR enhancements can improve the perceived usefulness and use of key functions. Successful scale-up and use of EHRs in small health facilities could improve clinical documentation, care, reporting, and disease surveillance in low- and middle-income countries.

(JMIR Med Inform 2022;10(5):e32305) doi:10.2196/32305

KEYWORDS

electronic health record; eHealth; HIV/AIDS; survey; Rwanda; implementation science

Introduction

Background

Effective and high-quality health care requires high-quality, timely health information—"Information is care" [1]. Scaling up effective care for millions of patients with HIV in resource-limited settings such as sub-Saharan Africa required the development of new paradigms for the collection, storage, viewing, and analysis of clinical data and health information [2]. Most health centers treating HIV started with only structured paper records. As the volume of patient data grew and in-country digital capacity improved, electronic tools were introduced. Many early electronic health records (EHRs) in resource-limited settings have been developed for HIV care, including those in Malawi [3], Kenya [4], and Haiti [5]. These projects demonstrated the feasibility of deploying health information systems, improvements in reporting to ministries of health (MoHs) and donors, and the ability to monitor the continuum of care. Furthermore, this initial evidence also suggested that the use of EHR systems for HIV, tuberculosis (TB) and multidrug-resistant TB treatment could improve the quality of care [2]. A critical challenge to improving the quality of care in low-income settings is the ability to achieve long-term, consistent EHR use at a large scale. To better understand the perceptions and clinical uses of EHR systems that support improved use and care in Rwanda, we conducted a quantitative user survey supplemented by free-text questions. For the purposes of this study, the terms *electronic health record* and *electronic medical record* are used interchangeably.

HIV Care in Rwanda

Rwanda is an East Central African country bordering Tanzania, Uganda, Burundi, and the Democratic Republic of the Congo. Rwanda had a per capita income of US $773 in 2018, up from US $241 in 2004 [6], and has made great progress in rebuilding its health care systems after the genocide against Tutsi in 1994. A major health challenge Rwanda has faced, along with neighboring countries in Africa, is the HIV epidemic. A 2018 to 2019 survey indicated that HIV prevalence among adults aged 15 to 49 years was 2.6% [7]. Great strides have been made in the treatment of patients who are HIV positive, including improvements in the prevention of mother-to-child transmission uptake, and reduction in the rate of loss to follow-up for patients receiving antiretroviral therapy (ART) in Rwanda. This is demonstrated by the near achievement in 2019 of the 2020 Joint United Nations Programme on HIV/AIDS 90-90-90 goal, with 84% of adults who were HIV positive knowing their status, 98% of those knowing their status on ART, and 90% of those on ART having a suppressed viral load [7]. From the beginning of the HIV treatment scale-up, the Government of Rwanda has emphasized care and prevention in rural areas as well as in urban settings; recruitment, training, and supervision of community health workers; and the use of health information systems. These information systems included national-level surveillance systems for HIV care [8,9], mobile health systems to support antenatal and primary care, and patient information or EHR systems mainly for supporting HIV care in health centers and hospitals. The 3 main EHR systems used have been OpenMRS (OpenMRS Inc) in health centers and 36 district hospitals offering HIV services, IQcare (International Quality Care, Palladium Inc) [10] in some health centers (now replaced by OpenMRS), and OpenClinic (OpenClinic GA) [11] in some hospitals. Since 2009, the MoH has moved to using OpenMRS for all HIV health centers and most hospitals in the country.

OpenMRS

OpenMRS is an open-source software platform for building EHRs, with a focus on health care needs in low- and middle-income countries (LMICs). Founded in 2004, the OpenMRS community set goals to create a public software platform to assist health care organizations worldwide in developing EHR systems that were adaptable to local needs, owned by local organizations, and programmed by local developers as much as possible [12] (Textbox 1).

Textbox 1. The OpenMRS electronic health record system.

OpenMRS has an unusual modular architecture allowing modules from the core development team to be mixed with modules from other developers to create flexible and updatable systems, with typical implementations using 35 to 45 modules. This ensures the core OpenMRS code is common to nearly all OpenMRS installations. Data are stored using a concept dictionary allowing flexibility in data capture and translation to other languages [12]. This approach also supports a range of standards for data storage and exchange with mappings available for a range of coding standards such as the International Classification of Diseases, 10th Revision, and Logical Observation Identifiers Names and Codes in the master Columbia International eHealth Laboratory concept dictionary.

Adapting OpenMRS to new uses typically requires technical expertise including Java programming if new modules are required. There were limitations to the older user interface used in this project (which has now been superseded), requiring care in developing clinical workflows. OpenMRS has been adapted to support a wide range of care including HIV, multidrug-resistant tuberculosis, primary care, emergency care, heart disease, oncology, and surgery. A Server Monitoring Tool module was developed to track system uptime and downtime, daily data entry rates, and completeness of key variables. The Server Monitoring Tool was used as part of the larger evaluation study in Rwanda.

OpenMRS was developed by a collaboration among the Academic Model Providing Access to Healthcare project in Kenya with the Regenstrief Institute in Indiana, United States; the Partners In Health Informatics team in Rwanda and Boston, Massachusetts, United States (HSF); and the informatics lead of the South African Medical Research Council (now CEO of Jembi Health Systems, Cape Town, South Africa—CS). Ongoing maintenance of the core OpenMRS platform is accomplished through the OpenMRS community—a worldwide network of volunteers with technology, health care, and international development expertise.

Initially, OpenMRS was used for HIV and TB treatment in outpatient settings, supporting projects funded by the US President's Emergency Plan for AIDS Relief and the Global Fund for AIDS, Tuberculosis, and Malaria. Currently, it covers a wide range of clinical areas. Partners In Health implemented and currently supports OpenMRS in 46 health centers and 3

hospitals in Rwanda covering HIV care, pediatrics, primary care, cardiology, and oncology.

Between 2009 and 2013, the Rwanda MoH deployed OpenMRS to >300 health centers providing HIV care throughout the country [13]. Before and during deployment, OpenMRS had to be customized to support the Rwanda MoH requirements. A dedicated 9-month course led by Partners In Health/Inshuti Mu Buzima trained programmers in enterprise Java and health information system design [14]. Several graduates were hired by the MoH and created custom OpenMRS modules for HIV and primary care using OpenMRS version 1.6 core code. This is the version of OpenMRS used in the control sites for this study. Unstable internet connectivity in rural Rwanda (similar to many low-income countries) required each site to run its own instance on a local server, requiring stable power and local technical support.

Impact of EHR Systems in Resource-Limited Countries

Over the last two decades, EHR systems have been implemented in a wide range of countries, including those with the lowest income levels. The scale-up of HIV care and transition from an emergency outbreak response to a lifelong chronic care model was a major driver for the expansion of EHR system use and the development of common, shared information system tools. Countries, including Rwanda, Kenya, Uganda, Mozambique, and Nigeria, have scaled up the use of OpenMRS EHR systems for HIV care to hundreds of their clinical sites. Other EHRs, including IQcare, have been widely used in countries such as Kenya [15].

The OpenMRS community has prioritized support for effective and safe clinical care as well as reporting and research. Smaller-scale studies have evaluated the impact of EHR system improvements on the aspects of clinical care, the systems *efficacy*. Were et al [16] studied the addition of alerts to printed patient summaries generated by OpenMRS on a range of clinical actions for the care of children who were HIV positive in Eldoret, Kenya. In a randomized controlled trial (RCT), they showed that health care workers receiving the summaries with alerts were 4 times more likely to carry out actions such as ordering CD4 counts (a T lymphocytes test) and polymerase chain reaction tests for HIV antigen. In a larger study, Oluoch et al [17] studied the impact of improved decision support tools implemented in an EHR in Kenya on the quality of HIV care. In a cluster RCT of 13 health centers and 41,062 patients, they showed that sites with the decision support tools were quicker and more effective in responding to HIV treatment failure [17]. Critical questions remain regarding the key factors that determine individual EHR use, facilitate scaling up to tens or hundreds of smaller health facilities, support long-term use, and influence the clinical impact of these systems in routine care—the *effectiveness* of EHRs in LMICs [18].

Methods

Overview

The aims of this study are to evaluate the following questions in a large number of health centers in Rwanda: (1) staff and stakeholder expectations and perceptions of health information system performance; (2) staff and stakeholder expectations and perceptions around effort expended to use health information systems; (3) infrastructural, organizational, and individual conditions that are barriers and facilitators to using such tools (including training and technical support); (4) staff perceptions of technology fatigue; and (5) any differences in the experiences of staff in intervention and control sites and between clinical and technical users.

The EHR Implementation Science Study

The focus of this manuscript is the electronic medical record (EMR) user survey component of a process evaluation, which is part of a larger, 3-part implementation science study on the use of an *enhanced* EHR to support HIV care in 56 randomly allocated health centers that commenced in July 2018. It included the evaluation of (1) EHR use, performance, and data quality; (2) the clinical impact in an RCT; and (3) the cost of development and implementation of the enhanced EHR functionality.

For enrollment in the overall study, first, the enhanced EHR package (Textbox 2) was piloted in 2 health centers in Kigali (Kicukiro Health Centre and Kagugu Health Centre), and improvements were made in response to the user experience and comments. Next, the following selection criteria were applied: (1) the presence of ≥3 computers, 1 printer, and a local area network; (2) active HIV case numbers between 50 and 700; and (3) successful installation of the Server Monitoring Tool (Textbox 1) and evidence of regular data entry by staff. Using these criteria, a total of 112 sites were selected to participate in the clustered RCT. These sites were a mix of urban and rural health centers and some district hospitals. Of the 112 sites, 56 (50%) were randomized into the intervention sites, which had the enhanced EHR installed on the servers between June 25 and July 5, 2018. All 56 sites had the alerts for delayed patient enrollment, 28 sites also had alerts for delayed viral load testing, and 14 had the alerts for evidence of treatment failure. For the analysis of the survey, sites with at least the top-level alerts (delayed HIV care registration) were classed as *intervention*. Health facility staff, including clinicians, data managers, local information technology (IT) staff, local clinic managers, and district IT specialists, in all 112 study sites were trained on general EHR use and data management. Additional training was provided for staff in the intervention sites on the enhanced EHR and equivalent training on the control EHR.

Textbox 2. The enhanced electronic health record (EHR) package.

The enhanced EHR package enhancements

- Upgraded OpenMRS software version (to v.1.11) and additions to the concept dictionary
- Improved workflow for registering and managing patients with HIV
- Improved ordering of laboratory analyses (HIV tests, CD4 counts, and viral loads)
- Upgraded clinician summaries of patients showing key clinical data and alerts and reminders designed to improve care
- Custom automatic reports to identify patients not receiving optimal care, implementing the same alerts and reminders
- Alerts and reports designed to identify patients with care delivery issues. These were chosen to reflect the needs identified by the Rwandan Ministry of Health and based on the 2016 World Health Organization guidelines for HIV care (WHO Consolidated Guidelines HIV 2016 [19]) and included the following:
  - Newly diagnosed patients with HIV who have not been enrolled in antiretroviral therapy within 2 weeks of diagnosis
  - Patients with 8 months of antiretroviral therapy who do not have a viral load test result in the EHR (6 months of care + 2 months for result to return and be entered in the EHR)
  - Patients who have an abnormal (elevated) viral load result and require assessment and management for treatment failure

Study Environment

The user survey was conducted at primary health care facilities, referred to here as health centers, offering HIV treatment services, located throughout Rwanda, approximately 5 months after the installation of the enhanced EHR.

Study Design

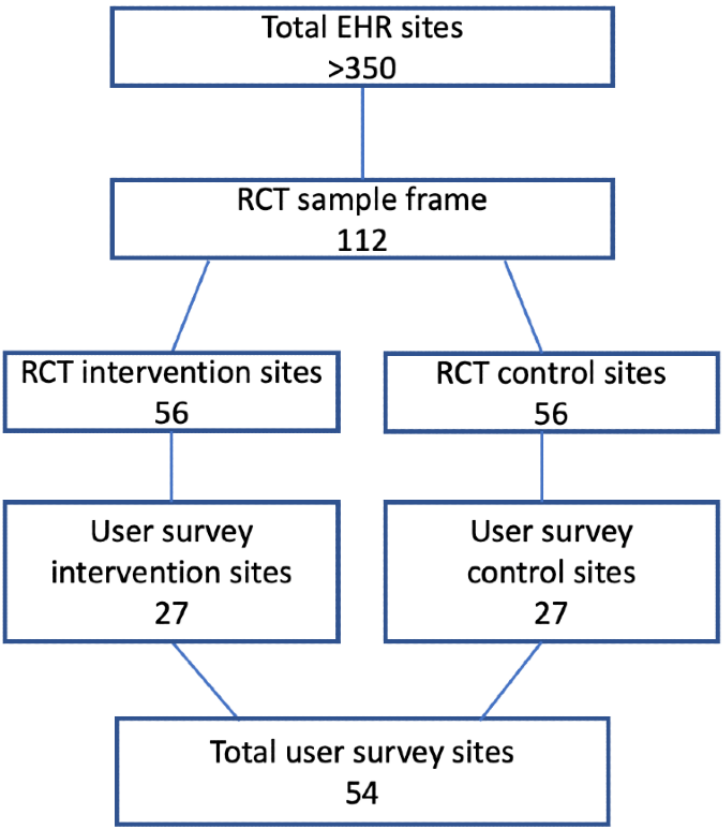
This study used a cross-sectional, key informant structured interview design within control and intervention sites. The data were collected through structured interviews to ensure high response rates and avoid technical limitations that may have impacted a web-based survey and biased results toward better-supported sites and users. The goal is to gain insights into the adoption, functionality, use, and perceptions of EHRs by clinical staff (nurses, physicians, and social workers) and technical staff (IT staff, data entry staff, and data managers) in health centers. Care of patients in smaller health facilities in East Africa, including those with HIV, is mostly carried out by nurses or junior clinician grades and rarely by physicians. The study questions were as follows: (1) whether the actions and perceptions of staff using the enhanced EHR intervention would be different from those using the control EHR and (2) whether clinicians have different experiences with the EHR than technical staff.

Sampling and Sample Size

This study drew from the sample frame of the clustered RCT implementation study. The RCT enrolled 112 health centers from >300 that use the OpenMRS EHR for HIV care. Of the 112 sites, 54 (48.2%) were randomly selected, including 27 (50%) from the enhanced EHR sites (intervention) and 27 (50%) control sites (Figure 1). Randomization was performed with R.

A total of 100 participants were approached for the structured interview, with the goal of 1 clinician (nurse or physician) and 1 data manager at each health center. If not available, other EHR users were recruited if possible.

Figure 1. Sample design (total electronic medical record sites include those managed by Partners In Health/Inshuti Mu Buzima; there were >300 active ministry of health–run sites). EHR: electronic health record; RCT: randomized controlled trial.



Data Collection Tool

The structured interview and observation (survey) tool included sections on demographics, experience with IT, EHR training received, frequency of EHR use for different tasks, overall ease of use, usefulness for specific tasks, technical and user support, and system stability and infrastructure issues. The survey tool included 5-point Likert scale quantitative close-ended questions and qualitative open-ended questions. It was adapted from a form originally used by Médecins Sans Frontières, piloted in 20 clinics in Rwanda in 2012 [13], and translated from English into Kinyarwanda.

Data Collection

The survey was conducted by 10 trained data collectors from the Rwanda School of Public Health. Survey responses were documented and recorded in a preprogrammed Android tablet using ODK [20]. Free-text comments were documented in Kinyarwanda, translated into English, and reviewed by bilingual research team members before analysis. Written informed consent was obtained, and participants' confidentiality was assured using a private interview room at each health center surveyed.

Data Analysis

Descriptive statistics were carried out using Excel (Microsoft Corp). JMP Statistical Software (SAS Institute) and Excel were used for the chi-square tests for the 5-point Likert scale responses. For the comparison of clinicians and technical users, all Likert scale questions were tested for significance. For the comparison of the intervention and control sites, the 2 groups of questions (12 and 18) most directly related to the technical improvements in the enhanced EHR were tested. *P* values were adjusted for multiple comparisons using the Benjamini and Hochberg method and R p.adjust [21]. Analyses were designed and carried out with assistance from a statistician and a data scientist at Brown University (see *Acknowledgments*).

Free-text comments, which were all short statements, were analyzed thematically using inductive coding by one author (HSF) and recoding by a second author (MM), with discrepancies resolved by discussion. Common concepts were described and rated based on the number of user responses matching each code.

Ethical Considerations

This study was approved by the following investigational review boards: Rwanda National Ethics Committee, Kigali (#913/RNAC/2016) and the University of Leeds School of Medicine Research Ethics Committee, Leeds, United Kingdom (MREC16-176). This study was reviewed in accordance with the US Centers for Disease Control and Prevention human research protection procedures (approval #CGH HSR 2014-270a) and determined to be research. However, investigators of the Centers for Disease Control and Prevention did not interact with human participants or have access to identifiable data during this research.

Results

Participant Characteristics

A total of 100 participants were approached for interview and consented to participate. Of these 100 participants, 90 (90% response rate) were available at the time of the visit by study staff, with 44 (49%) from the intervention sites. The participants had a mean age of 35.9 (SD 6.2; range 23-58) years, a mean of 7.5 (SD 2.0; range 0-38) years working at the health centers, and 47% (42/90) were female. Their educational attainment was reported as *some secondary schooling* (8/90, 9%), *completed secondary schooling* (18/90 20%), and *postsecondary schooling* (64/90, 71%). Their occupations were nurse (41/90, 46%), physician (1/90, 1%), social worker (2/90, 2%), data manager (42/90, 47%), IT officer (1/90, 1%), and data entry staff (3/90, 3%). Respondents had a mean of 3.3 (SD 2.0) years of experience with the EHR and a mean of 1.9 (SD 1.4) trainings and 9.7 (SD 7.9) training days.

General Use of Technology

A large majority of respondents used mobile phones, with 82% (74/90) using them "about half the time," "most of the time," or "all of the time" for texting, and 82% (74/90) similarly for mobile data. Computer use outside of work was reported by 31% (28/90) and internet use by 49% (44/90). Clinicians (physicians, nurses, and social workers) reported significantly less use than technical staff (IT officers, data managers, and data entry staff; P=.009 and P=.04, respectively).

Training on EHR

Respondents agreed or strongly agreed that their training on the EHR was effective (82/84, 98%), and they were confident in using the EHR (81/87, 93%). However, 77% (66/86) of respondents disagreed or strongly disagreed with the statement "I am generally not concerned making errors in EHR." There were no statistically significant differences in responses on training between clinicians and technical staff. However, in free-text comments, 81% (73/90) of respondents requested more training. These requests included refreshers, training on new modules or updates, and more practical hands-on training. There were also requests for training in reports and data analysis. Mentorship, supportive supervision, or more technical backup were requested by many respondents.

Use of EHR Functions

Tables 1-7 show and summarize the results for the following question: "Please indicate how often you use the EMR to assist you with the following tasks." Combining the categories most of the occasions and always/almost always, the percentages for common tasks were 56% (49/88) for creating new patient records, 63% (56/89) for updating existing patient records, 40% (36/89) for generating patient summaries before visits, 48% (43/89) for reviewing previous patient encounters, 30% (21/69) for ordering laboratory analyses, 43% (36/83) for viewing laboratory results, 33% (25/75) for following test results over time, 45% (38/85) for generating automatic reports, 45% (38/85) for generating ad hoc reports (eg, quarterly or TracNET reports), and 49% (41/84) for referring patients to another health facility. The results were 22% (18/82) for generating consult sheets and 16% (14/85) for generating clinician summaries.

Table 1. Frequency of survey responses for Likert scale data: question 6 (n=90).

Question 6. How often do you do the following activities?	1—never/almost never, n (%)	2—seldom, n (%)	3—about half the occasions, n (%)	4—most of the occasions, n (%)	5—always/almost always, n (%)	Top 2 groups, n (%)
Use a mobile phone to send text messages	5 (6)	3 (3)	8 (9)	40 (44)	34 (38)	74 (82)
Use a mobile phone to access email, internet, WhatsApp, or Facebook	5 (6)	2 (2)	9 (10)	44 (49)	30 (33)	74 (82)
Use a computer outside of work	28 (31)	5 (6)	29 (32)	16 (18)	12 (13)	28 (31)
Access the internet to check email, go to websites, or any other internet activities	13 (14)	13 (14)	20 (22)	27 (30)	17 (19)	44 (49)

Table 2. Frequency of survey responses for Likert scale data: question 10.

Question 10. Training	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)	Top 2 groups, n (%)
The training I received relating to the EMR[a] was effective (n=84)	1 (1)	1 (1)	0 (0)	21 (25)	61 (73)	82 (98)
In general I am not concerned about making errors in the EMR (n=86)	36 (42)	30 (35)	3 (3)	13 (15)	4 (5)	17 (20)
I am confident using the EMR (n=87)	1 (1)	4 (5)	1 (1)	42 (48)	39 (45)	81 (93)

[a]EMR: electronic medical record.

Table 3. Frequency of survey responses for Likert scale data: questions 12 to 14.

Question	Never/almost never, n (%)	Seldom, n (%)	About half of the occasions, n (%)	Most of the occasions, n (%)	Always/almost always, n (%)	Top 2 groups, n (%)
12. Please indicate how often you use the electronic medical record to assist you with the following tasks.						
Creating new patient records (n=88)	2 (2)	14 (16)	23 (26)	17 (19)	32 (36)	49 (56)
Updating existing patient records (n=89)	3 (3)	10 (11)	20 (22)	26 (29)	30 (34)	56 (63)
Generating patient summaries before visits (n=89)	11 (12)	14 (16)	28 (31)	21 (24)	15 (17)	36 (40)
Reviewing previous patient encounters (n=89)	6 (7)	13 (15)	27 (30)	21 (24)	22 (25)	43 (48)
Ordering laboratory analyses (n=69)	32 (46)	5 (7)	11 (16)	12 (17)	9 (13)	21 (30)
Viewing laboratory results (n=83)	22 (27)	7 (8)	18 (22)	19 (23)	17 (20)	36 (43)
Following test results over time (n=75)	32 (43)	4 (5)	14 (19)	15 (20)	10 (13)	25 (33)
Ordering medicine (n=65)	50 (77)	3 (5)	5 (8)	2 (3)	5 (8)	7 (11)
Generating pharmacy reports (n=79)	54 (68)	4 (5)	7 (9)	7 (9)	7 (9)	14 (18)
Generating automatic reports (n=85)	31 (36)	8 (9)	8 (9)	14 (16)	24 (28)	38 (45)
Generating ad hoc reports (n=85)	34 (40)	2 (2)	11 (13)	13 (15)	25 (29)	38 (45)
Generating consult sheets (n=82)	47 (57)	5 (6)	12 (15)	12 (15)	6 (7)	18 (22)
Generating clinician summaries (n=85)	48 (56)	8 (9)	15 (18)	8 (9)	6 (7)	14 (16)
Referring patients to another health center (n=84)	20 (24)	6 (7)	17 (20)	14 (17)	27 (32)	41 (49)
13. All considered, how often do you use the electronic medical record as an information source in your clinical work? (n=89)	13 (15)	10 (11)	29 (33)	38 (40)	1 (1)	37 (42)
14. All considered, how often do you use paper-based medical records as an information source in your clinical work? (n=89)	4(4)	4 (4)	13 (15)	61 (69)	7 (8)	68 (76)

Table 4. Frequency of survey responses for Likert scale data: question 16.

Question 16. Please tell us the degree to which you agree or disagree with the following statements about the EMR[a].	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)	Top 2 groups, n (%)
I am able to find where to document care (n=84)	6 (7)	3 (4)	4 (5)	49 (58)	22 (26)	71 (85)
In general it is easy to correct errors in EMR (n=89)	4 (4)	19 (21)	2 (2)	44 (49)	20 (22)	64 (72)
In general the screen display is easy to read (n=89)	1 (1)	2 (2)	1 (1)	38 (43)	47 (53)	85 (96)
The content is laid out in an understandable way (n=89)	1 (1)	6 (7)	5 (6)	53 (60)	24 (27)	77 (87)
It is easy to retrieve patient records in the EMR (n=89)	1 (1)	4 (4)	1 (1)	44 (49)	39 (44)	83 (93)

[a]EMR: electronic medical record.

Table 5. Frequency of survey responses for Likert scale data: question 18.

Question 18. Please tell us the degree to which you agree or disagree with the following statements about the EMR[a].	Never/almost never, n (%)	Seldom, n (%)	About half of the occasions, n (%)	Most of the occasions, n (%)	Always/almost always, n (%)	Top 2 groups, n (%)
The EMR provides useful alerts, reminders (n=82)	5 (6)	7 (9)	6 (7)	36 (44)	28 (34)	64 (78)
The EMR makes it easier to manage patients (n=90)	0 (0)	2 (2)	1 (1)	37 (41)	50 (56)	87 (97)
The EMR easier to make informed decisions (n=90)	1 (1)	3 (3)	1 (1)	40 (44)	45 (50)	85 (94)
The EMR makes it easier exchange patient information with other health care providers (n=90)	0 (0)	21 (23)	5 (6)	34 (38)	30 (33)	64 (71)
The EMR is worth the time and energy to use (n=90)	0 (0)	1 (1)	0 (0)	44 (49)	45 (50)	89 (99)
The quality of information has improved due to the EMR (n=90)	0 (0)	3 (3)	4 (4)	50 (56)	33 (37)	83 (92)

[a]EMR: electronic medical record.

Table 6. Frequency of survey responses for Likert scale data: question 20.

Question 20. Please tell us the degree to which you agree or disagree with the following statements about the EMR[a].	Strongly disagree, n (%)	Disagree, n (%)	Neutral, n (%)	Agree, n (%)	Strongly agree, n (%)	Top 2 groups, n (%)
It is easy to report problems with the EMR (n=89)	7 (8)	17 (19)	2 (2)	45 (51)	18 (20)	63 (71)
I get feedback when I report errors or problems with the EMR (n=89)	7 (8)	23 (26)	6 (7)	45 (51)	8 (9)	53 (60)
Effective help is available when I experience problems with the EMR (n=89)	9 (10)	28 (31)	3 (3)	40 (45)	9 (10)	49 (55)
I use the EMR because of the proportion of coworkers who use it (n=86)	11 (13)	37 (43)	3 (3)	27 (31)	8 (9)	35 (41)
My supervisor is very supportive of use of the EMR for my job (n=89)	7 (8)	9 (10)	6 (7)	39 (44)	28 (31)	67 (75)
In general, the Ministry of Health has supported the use of the EMR (n=90)	1 (1)	2 (2)	3 (3)	49 (54)	35 (39)	84 (93)

[a]EMR: electronic medical record.

Table 7. Frequency of survey responses for Likert scale data: question 22.

Question 22. Indicate how often you experience the following:	Never/almost never, n (%)	Seldom, n (%)	About half of the occasions, n (%)	Most of the occasions, n (%)	Always/almost always, n (%)	Top 2 groups, n (%)
How often can you count on EMR[a] to be up and available? (n=88)	2 (2)	4 (5)	16 (18)	29 (33)	37 (42)	66 (75)
How often is grid electricity present? (n=90)	3 (3)	3 (3)	10 (11)	38 (42)	36 (40)	74 (82)
How often is the backup generator available? (n=88)	52 (59)	3 (3)	3 (3)	4 (5)	26 (30)	30 (34)
How often is there internet? (n=89)	12 (13)	2 (2)	19 (21)	20 (22)	36 (40)	56 (63)
How often is there cellular network coverage? (n=85)	27 (32)	6 (7)	11 (13)	13 (15)	28 (33)	41 (48)
How often is a computer available when you need to use the EHR[b]? (n=89)	4 (4)	2 (2)	6 (7)	12 (13)	65 (73)	77 (87)
How often is the EHR very slow? (reverse scale; n=88)	31 (35)	15 (17)	29 (33)	9 (10)	4 (5)	13 (15)

[a]EMR: electronic medical record.

[b]EHR: electronic health record.

Staff in intervention sites were significantly more likely to use the EHR for "Updating existing patient records" (P=.04), "Generating patient summaries before visits" (P<.001), "Viewing laboratory results" (P=.04), and "Generating clinician summaries" (ie, on-screen summaries; P=.01). Clinician responses indicated that they carried out the following tasks significantly less frequently than technical staff: "Creating new patient records" (P=.02) and "Updating existing patient records" (P=.04).

A total of 42% (37/89) of respondents stated that they used the EHR always/almost always or most of the time, as opposed to 76% (68/89) for the paper records. They *agreed or strongly agreed* >85% (71/84) of the time (Tables 1-7) with the following statements about the EMR: "I can find where to document care," "The screen displays are easy to read," "Content lay out is understandable," and "It is easy to retrieve records in EHR." For the statement "It is easy to correct errors in EHR," agreement was 72% (64/89).

Respondents *agreed or strongly agreed* >90% (81/90) of the time that "the EHR makes it easier to manage patients' medical file and patient's medical follow up," "the EHR makes it easier to make informed decisions," "the EHR is worth the time and energy to use," and "quality of information has improved due to the EHR." For the statement "the EHR makes it easier to exchange patient information with other health care providers," agreement was 71% (64/90). For the statement "the EHR provides useful alerts and reminders," agreement was 78% (64/82) with significantly stronger agreement in the intervention sites (P=.01).

Answers to questions on technical and user support received mixed responses. Respondents *agreed or strongly agreed* with these questions with the following scores: "It is easy to report problems with the EHR," 71% (63/89); "I get feedback when I report errors or problems," 60% (53/89); "Effective help is available with the EHR," 55% (49/89); "I use EHR because of the proportion of coworkers who use it," 41% (35/86); "My supervisor is very supportive of EHR use on the job," 75% (67/89); and "In general, the MOH supported the use of EHR," 93% (84/90).

Infrastructure

Infrastructure problems were a significant issue (Tables 1-7). The following were stated to be available *always/almost always* or *most of the occasions*: a computer when you need the EHR (77/89, 87%), grid power (74/90, 82%), wired internet connectivity (56/89, 63%), cellular internet (41/85, 48%), and a backup generator (30/88, 34%). For the question "How often can you count on EHR to be up and available?" response was 75% (66/88), with 18% (16/88) saying it was available about half the time and 7% (6/88) almost never.

Table 8 shows the analysis of free-text comments. The most frequent responses to the question "What are three functions you like about the electronic medical record?" were "to get client data easily and/or quickly" (62/90, 69%), "it helps to generate reliable reports in a short time" (39/90, 43%), "it stores client information safely and/or securely" (31/90, 34%), and "it helps to monitor clients on a daily basis" (20/90, 22%). In response to the question "What are three functions you do not like about the electronic medical record?" most frequent comments were "often unstable or blocked" (20/90, 22%), "hard to correct errors or unsubscribe patients" (11/90, 12%), "cannot work with OpenMRS outside the health facility/not online" (9/90, 10%), and "poor internet" (6/90, 7%).

Table 8. Responses to free-text questions on user likes and dislikes (n=90).

Question, themes, and example comments	Value, n (%)
What are 3 functions you like about the electronic medical record?	
Supports accessible and safe patient record keeping	
Helps users to get client data easily and/or quickly	62 (69)
Stores client information safely and/or securely	31 (34)
Supports patient care by providing needed information on the patient	
Provides alerts	6 (6)
Makes managing patient data easier	
"Simplifies my daily work"	14 (16)
Helps to generate reports	
Support generation of reports reliably and in a short time	39 (43)
Example comments	
"It provides an alert regarding viral loads, CD4." (intervention site)	N/A[a]
"When it is well manipulated can reduce workload in the service."	N/A
"It indicates missing information in the client's file."	N/A
"To identify client that do not respect their appointment."	N/A
"Number of lost follow up."	N/A
What are 3 functions you do not like about the electronic medical record?	
System stability or unavailability	
Often unstable or blocked	20 (22)
Lack of technical support	7 (8)
Poor internet connection	6 (7)
Lack of updates for key functionality or metadata	
Lack of drugs listed in formulary	3 (3)
Lack of connectivity beyond individual health facilities	
Cannot work with OpenMRS outside the health facility/not on the internet	9 (10)
Unable to track patient transfers	4 (4)
Error correction/editing	
Hard to correct errors	7 (8)
Cannot unsubscribe patients	2 (2)
Example comments	
"There are few nurses that use OpenMRS efficiently."	N/A
"You cannot use OpenMRS out of working site."	N/A
"I like OpenMRS but this new version there some information that cannot provide."	N/A
"I like OpenMRS but this new version there some information that cannot provide."	N/A
"Blockage of OpenMRS affects my daily performance."	N/A

[a]N/A: not applicable.

Discussion

Principal Findings

Overall, the results suggest that most users of OpenMRS at Rwanda MoH health centers perceive the EHR as a valuable tool for patient care and reporting activities. The responses showed a high level of EHR use and acceptability across most health centers despite the challenges of implementing EHR systems in these environments. This finding provides foundational evidence to implementers who have an urgent need to understand how well EHRs can be scaled up to hundreds or thousands of health facilities (addressing objectives on performance and scalability). An unusual feature of the Rwanda

OpenMRS implementation is the long interval since the original deployment. Some MoH health centers have used the EHR continuously for 8 or 9 years, with no major upgrades in control sites for >5 years. Therefore, this study allows assessment of the long-term performance of the EHR by typical users (objective on sustainability). Such data are not available from any existing studies that we are aware of, which have mostly focused on larger hospitals in more controlled settings with better infrastructure [11,22] or small numbers of test sites.

Responses to the 2 groups of questions most relevant to the features of the enhanced EHR package showed, in the intervention sites, more frequent use of core clinical tools, including updating records, using patient summaries, and viewing laboratory results. Significantly more respondents in the intervention sites agreed that "The electronic medical record provides useful alerts and reminders," indicating support for more advanced EHR features added in the enhanced EHR.

There were some differences in the level of EHR use between clinicians and technical staff, including core clinical activities such as creating and updating records. Clinicians, as expected, had less technical experience and were significantly less likely to use computers outside work or access the internet for a range of applications. These findings indicate the need for further improvements in usability and workflow and in both IT and EHR training for clinicians.

It is important to note that recent versions of OpenMRS have greatly improved user interfaces and general functionality [23,24] and are expected to have significantly higher scores for usability and overall satisfaction. An up-to-date version of OpenMRS was implemented in Rwandan district hospitals in 2020/2021.

Limitations

This survey was conducted through structured interviews with all participants. The less confidential nature of interviews compared with a web-based survey may have increased *desirability bias* as staff were aware that the study was endorsed by the MoH. There was a strong positive response on the question of MoH support and on statements that the effort to enter data and use the EHR was worthwhile. However, on other questions such as infrastructure, including power and internet connectivity, and availability of technical support, participants were more mixed in their responses, and for the question "I am generally not concerned making errors in EHR" they were clearly prepared to admit that there were problems. Many made clear that they had challenges with using the EHR, and clinicians would appear to rely on data managers and other technical staff to assist with many activities. Free-text comments provided critical insights into the actual experiences of staff, along with many other issues related to usability, use, and the need for training. The lack of significant differences in the experiences of the clinicians and technical staff regarding many questions may be partly due to the survey not being powered to show small differences between these groups. Another limitation was that the 112 sites selected for the broader study had better hardware and evidence of more consistent data entry than the others; therefore, EMR implementations described here may perform better than the full set of EMR sites in Rwanda.

Comparison With Previous Work

Previous studies of EHR users in LMICs have identified a range of experiences. Ojo [25] used the Delone McLean Information Success Model in a study of EHR users in hospitals in Nigeria and showed that system quality and use were the most important in determining EHR success [25]. Tilahun and Fritz [26] conducted a similar study on the experience of users with an EHR in hospitals in Ethiopia. Compared with the survey in this study, they showed high levels of dissatisfaction with the EHR and low use levels owing to poor service quality (power infrastructure, user support, training, and lack of computers in the wards) and the need for double entry of data into the EHR and paper records (also a problem in Rwanda) [26]. A survey of the OpenClinic EHR users at the Kigali University Teaching Hospital in Rwanda showed strongly positive user comments on satisfaction and perception of data quality and usability compared with paper records [11].

Conclusions

This survey provides evidence that EHR systems have become an accepted component of HIV care delivery in Rwanda. Staff were generally supportive of the system, although most wanted further training, technical support, and better power and network infrastructure. Staff at intervention sites were more likely to use or have positive experiences of key functionality that was improved in the enhanced EHR. As this survey is part of a larger evaluation study, the responses will be compared with results from key informant interviews, the costing and data quality studies, monitoring of server performance and use, and clinical impact in the cluster RCT. Further surveys are planned for other large-scale rollouts of OpenMRS in low-income settings, building on the survey form and findings in this study. The results are likely to be generalized to similar EHR systems in low-income settings if they are well tailored to the clinical needs and workflow. They are also highly relevant to the critical need for systems to support accurate, timely, and analyzable primary care data on patients in remote and very underserved clinics in low-income countries, replacing basic tools such as paper registers. This should improve the clinical documentation, care, reporting, and tracking of disease outbreaks, including COVID-19.

Data Availability

The data underlying this paper cannot be shared publicly because of the need for privacy of the individuals who participated in the study. The data will be shared upon reasonable request with the corresponding author.

Acknowledgments

The authors thank the clinic staff at the study sites in Rwanda for their participation in the survey, the staff of the Rwanda Biomedical Center and the Ministry of Health, and the informatics team at Partners In Health in Rwanda for technical assistance and advice. The authors thank Dr Tao Liu and Dr Ian Bacher for assistance with statistical analysis and Dr Eric Green (Duke University) for assistance with the RCT design and randomization.

This research was supported by the President's Emergency Plan for AIDS Relief through the Centers for Disease Control and Prevention under program grant U01 GH000782-01 awarded to the National University of Rwanda. Additional funding was provided by Brown University, Providence, Rhode Island, United States. A pilot study of the user survey was funded by the Rockefeller Foundation (grant 2010 THS 312). HSF was a Marie Skłodowska-Curie Fellow 2015-2017, funded by the European Union's Horizon 2020 research and innovation program under grant agreement 661289, "Global eHealth."

The development of OpenMRS was funded by several sources, including the Rockefeller Foundation, US Centers for Disease Control and Prevention, World Health Organization, International Development Research Centre (Ottawa, Ontario, Canada), and Partners In Health.

Authors' Contributions

HSF, MM, JR, XS, WN, CS, JC, and AU designed the study. All authors assisted in the collection and/or analysis of the data. HSF, MM, JR, XS, CS, JC, and AU drafted the manuscript. All the authors critically reviewed the manuscript for content and signed the final version.

Conflicts of Interest

HSF and CS are cofounders of the OpenMRS electronic health record project. JR and XS worked for the Centers for Disease Control and Prevention Division of Global HIV and Tuberculosis, which funded this study. None of these individuals received compensation for their time from the Centers for Disease Control and Prevention implementation science grant. HSF received travel and accommodation support for study meetings and workshops from the grant. Jembi Health Systems, whose director is CS, received funds from a grant for software development work on OpenMRS. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the funding agencies.

Multimedia Appendix 1
Survey form in English and Kinyarwanda.
[DOCX File , 88 KB - medinform_v10i5e32305_app1.docx]

References

1. Berwick D. Escape Fire: Designs for the Future of Health Care. San Francisco, CA: John Wiley & Sons, Inc; 2004.
2. Fraser HS, Allen C, Bailey C, Douglas G, Shin S, Blaya J. Information systems for patient follow-up and chronic management of HIV and tuberculosis: a life-saving technology in resource-poor areas. J Med Internet Res 2007 Oct 22;9(4):e29 [FREE Full text] [doi: 10.2196/jmir.9.4.e29] [Medline: 17951213]
3. Douglas G, Gadabu O, Joukes S, Mumba S, McKay MV, Ben-Smith A, et al. Using touchscreen electronic medical record systems to support and monitor national scale-up of antiretroviral therapy in Malawi. PLoS Med 2010 Aug 10;7(8):e1000319 [FREE Full text] [doi: 10.1371/journal.pmed.1000319] [Medline: 20711476]
4. Rotich JK, Hannan TJ, Smith FE, Bii J, Odero WW, Vu N, et al. Installing and implementing a computer-based patient record system in sub-Saharan Africa: the Mosoriot Medical Record System. J Am Med Inform Assoc 2003;10(4):295-303 [FREE Full text] [doi: 10.1197/jamia.M1301] [Medline: 12668697]
5. Fraser H, Jazayeri D, Nevil P, Karacaoglu Y, Farmer P, Lyon E, et al. An information system and medical record to support HIV treatment in rural Haiti. BMJ 2004 Nov 13;329(7475):1142-1146 [FREE Full text] [doi: 10.1136/bmj.329.7475.1142] [Medline: 15539669]
6. GDP per capita (current US$) - Rwanda. The World Bank. URL: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=RW [accessed 2022-04-20]
7. Rwanda Population-based HIV Impact Assessment. ICAP University of Columbia. URL: https://phia.icap.columbia.edu/wp-content/uploads/2019/10/RPHIA-Summary-Sheet_Oct-2019.pdf [accessed 2022-04-20]
8. Kayumba K, Nsanzimana S, Binagwaho A, Mugwaneza P, Rusine J, Remera E, et al. TRACnet internet and short message service technology improves time to antiretroviral therapy initiation among HIV-infected infants in Rwanda. Pediatr Infect Dis J 2016 Jul;35(7):767-771 [FREE Full text] [doi: 10.1097/INF.0000000000001153] [Medline: 27031258]
9. DHIS2 Documentation homepage. DHIS2 Documentation. URL: https://docs.dhis2.org/en/home.html [accessed 2022-04-20]
10. Ali K, Mysha S, Gikandi N, Joshua O, Donna M. Bending the Cost Curve in Implementing Electronic Medical Record Systems: Lessons Learnt from Kenya. Calgary, AB: ACTA Press; 2014.

11. Uwambaye P, Njunwa K, Nuhu A, Kumurenzi A, Isyagi M, Murererehe J, et al. Health care consumer's perception of the Electronic Medical Record (EMR) system within a referral hospital in Kigali, Rwanda. Rwanda J 2017;4(1):48-53. [doi: 10.4314/rj.v4i1.7F]
12. Mamlin BW, Biondich PG, Wolfe BA, Fraser H, Jazayeri D, Allen C, et al. Cooking up an open source EMR for developing countries: OpenMRS - a recipe for successful collaboration. AMIA Annu Symp Proc 2006:529-533 [FREE Full text] [Medline: 17238397]
13. National roll out of the Rwanda OpenMRS electronic medical record to improve healthcare delivery. AMIA 2013. URL: https://dblp.org/pid/148/5735.html [accessed 2022-04-20]
14. Seymour RP, Tang A, DeRiggi J, Munyaburanga C, Cuckovitch R, Nyirishema P, et al. Training software developers for electronic medical records in Rwanda. Stud Health Technol Inform 2010;160(Pt 1):585-589. [Medline: 20841754]
15. Muinga N, Magare S, Monda J, Kamau O, Houston S, Fraser H, et al. Implementing an open source electronic health record system in Kenyan health care facilities: case study. JMIR Med Inform 2018 Apr 18;6(2):e22 [FREE Full text] [doi: 10.2196/medinform.8403] [Medline: 29669709]
16. Were MC, Nyandiko WM, Huang KT, Slaven JE, Shen C, Tierney WM, et al. Computer-generated reminders and quality of pediatric HIV care in a resource-limited setting. Pediatrics 2013 Mar;131(3):e789-e796. [doi: 10.1542/peds.2012-2072] [Medline: 23439898]
17. Oluoch T, Katana A, Kwaro D, Santas X, Langat P, Mwalili S, et al. Effect of a clinical decision support system on early action on immunological treatment failure in patients with HIV in Kenya: a cluster randomised controlled trial. Lancet HIV 2016 Feb;3(2):e76-e84 [FREE Full text] [doi: 10.1016/S2352-3018(15)00242-8] [Medline: 26847229]
18. Puttkammer N, Zeliadt S, Balan J, Baseman J, Destiné R, Domerçant JW, et al. Development of an electronic medical record based alert for risk of HIV treatment failure in a low-resource setting. PLoS One 2014;9(11):e112261 [FREE Full text] [doi: 10.1371/journal.pone.0112261] [Medline: 25390044]
19. Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection (2016). Geneva: World Health Organization; 2016.
20. Collect data anywhere. Open Data Kit. URL: https://getodk.org [accessed 2022-04-20]
21. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
22. Fritz F, Tilahun B, Dugas M. Success criteria for electronic medical record implementations in low-resource settings: a systematic review. J Am Med Inform Assoc 2015 Mar;22(2):479-488. [doi: 10.1093/jamia/ocu038] [Medline: 25769683]
23. Bahmni Wiki. Bahmni. URL: https://bahmni.atlassian.net/wiki/spaces/BAH/overview [accessed 2022-04-20]
24. OpenMRS Medical Record System. OpenMRS. URL: https://www.openmrs.org [accessed 2022-04-20]
25. Ojo AI. Validation of the DeLone and McLean information systems success model. Healthc Inform Res 2017 Jan;23(1):60-66 [FREE Full text] [doi: 10.4258/hir.2017.23.1.60] [Medline: 28261532]
26. Tilahun B, Fritz F. Comprehensive evaluation of electronic medical record system use and user satisfaction at five low-resource setting hospitals in Ethiopia. JMIR Med Inform 2015 May 25;3(2):e22 [FREE Full text] [doi: 10.2196/medinform.4106] [Medline: 26007237]

Abbreviations

EHR: electronic health record
EMR: electronic medical record
IT: information technology
LMIC: low- and middle-income country
MoH: ministry of health
RCT: randomized controlled trial

Edited by C Lovis; submitted 22.07.21; peer-reviewed by M Randriambelonoro, Y Chu, FJ Sánchez-Laguna; comments to author 14.11.21; revised version received 08.01.22; accepted 31.01.22; published 03.05.22.

Please cite as:
Fraser HSF, Mugisha M, Remera E, Ngenzi JL, Richards J, Santas X, Naidoo W, Seebregts C, Condo J, Umubyeyi A
User Perceptions and Use of an Enhanced Electronic Health Record in Rwanda With and Without Clinical Alerts: Cross-sectional Survey
JMIR Med Inform 2022;10(5):e32305
URL: https://medinform.jmir.org/2022/5/e32305
doi:10.2196/32305
PMID:35503526

Original Paper

Clustering Diagnoses From 58 Million Patient Visits in Finland Between 2015 and 2018

Pasi Fränti[1], PhD; Sami Sieranoja[1], PhD; Katja Wikström[2,3], PhD; Tiina Laatikainen[2,3], PhD

[1]Machine Learning Group, School of Computing, University of Eastern Finland, Joensuu, Finland

[2]Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland

[3]The Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland

Corresponding Author:
Sami Sieranoja, PhD
Machine Learning Group
School of Computing
University of Eastern Finland
Box 111
Joensuu, 80101
Finland
Phone: 358 405929966
Email: samisi@cs.uef.fi

Abstract

Background: Multiple chronic diseases in patients are a major burden on the health service system. Currently, diseases are mostly treated separately without paying sufficient attention to their relationships, which results in the fragmentation of the care process. The better integration of services can lead to the more effective organization of the overall health care system.

Objective: This study aimed to analyze the connections between diseases based on their co-occurrences to support decision-makers in better organizing health care services.

Methods: We performed a cluster analysis of diagnoses by using data from the Finnish Health Care Registers for primary and specialized health care visits and inpatient care. The target population of this study comprised those 3.8 million individuals (3,835,531/5,487,308, 69.90% of the whole population) aged ≥18 years who used health care services from the years 2015 to 2018. They had a total of 58 million visits. Clustering was performed based on the co-occurrence of diagnoses. The more the same pair of diagnoses appeared in the records of the same patients, the more the diagnoses correlated with each other. On the basis of the co-occurrences, we calculated the relative risk of each pair of diagnoses and clustered the data by using a graph-based clustering algorithm called the M-algorithm—a variant of k-means.

Results: The results revealed multimorbidity clusters, of which some were expected (eg, one representing hypertensive and cardiovascular diseases). Other clusters were more unexpected, such as the cluster containing lower respiratory tract diseases and systemic connective tissue disorders. The annual cost of all clusters was €10.0 billion, and the costliest cluster was cardiovascular and metabolic problems, costing €2.3 billion.

Conclusions: The method and the achieved results provide new insights into identifying key multimorbidity groups, especially those resulting in burden and costs in health care services.

(JMIR Med Inform 2022;10(5):e35422) doi:10.2196/35422

KEYWORDS

multimorbidity; cluster analysis; disease co-occurrence; multimorbidity network; health care data analysis; graph clustering; k-means; data analysis; cluster; machine learning; comorbidity; register; big data; Finland; Europe; health record

Introduction

Multimorbidity

Multiple chronic diseases in patients are a major burden to the health service system in terms of both service use and costs [1]. In many service systems, diseases are mostly treated separately without paying sufficient attention to their relationships, which results in the fragmentation of the care process. Better integration of services can lead to a more effective organization of the overall health care system. To support this, we analyzed

the connections between diseases based on their co-occurrence and performed a clustering analysis to identify multimorbidity patterns.

Multimorbidity is often defined as the coexistence of ≥2 chronic conditions within a patient [2,3]; however, the number of medical conditions included in this definition ranges widely [4]. Systematic reviews have shown that multimorbidity reduces self-rated health, quality of life, and functional ability and increases the risk of premature death, hospitalization, and use of health services, causing a substantial economic burden for societies and health care systems [5]. Wang et al [6] reported that multimorbidity cases, defined as patients with ≥2 chronic conditions, have 2 to 16 times higher costs than nonmultimorbidity cases. Brettschneider et al [7] analyzed the impact of 45 conditions on health-related quality of life. The authors measured multimorbidity using a weighted count score and assessed its association with decreases in the health-related quality of life. The strongest impact was observed in Parkinson disease, depression, and obesity.

An active research area is the measurement of the severity of multimorbidity. Stirland et al [8] reviewed 35 multimorbidity measures. Most measures (25 of 35) in their review were based on simple (weighted or unweighted) counts of diseases; some measures (4 of 35) used drug counts, and some (5 of 35) were based on expert-generated grouping of diagnoses, mainly based on frequencies. Such measures have been used to assess mortality, health care use, cost, and quality of life.

Diagnosis Groups

The number of possible multimorbidities is too large for human analysts to examine them individually. In the case of only 205 diagnoses, there are 20,910 different pairs of diagnoses. It is easier to analyze their connections by first dividing the diagnoses into smaller groups that contain related diagnoses and then examining only the connections between diagnoses within each group. This effectively removes less relevant multimorbidities from the data and allows us to show the connections in small groups that are easy to analyze.

Diagnosis groups can also predict future costs for a patient. Farley [9] discovered that simply counting the number of diagnosis clusters to which a patient belongs is a good predictor of high costs in the future. When combined with other measures such as the number of prescriptions, it outperformed more complex comorbidity indices such as the Charlson, Elixhauser, and RxRisk-V indices [9].

Diagnosis groups were previously created manually by experts by joining diagnoses of clinical similarity. Travers et al [10] studied how well the 4 groupings covered emergency medicine. The authors discovered that the Agency for Healthcare Research and Quality grouping for inpatient care provides the best coverage (99%), whereas the National Center for Health Statistics vital statistics grouping covers only 88%. They also criticized that most clusters (76%) were small, and there were large clusters containing dissimilar conditions. Open questions include how to evaluate a cluster system and determine its clinical relevance. Travers et al [10] further argued that a good clustering system should collapse the individual International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes into clinically meaningful clusters.

The number of groups was also problematic. Schneeweiss et al [11] argued that 367 clusters are too many for comparative analysis, whereas 17 clusters are too broad for this purpose. The authors reduced the number of International Classification of Diseases (ICD) categories to 110 diagnosis clusters by cross-tabulation between the ICD-9-CM and International Classification of Health Problems in Primary Care–2 classifications, covering approximately 90% of all diagnoses of their records made by family physicians.

Clustering to Detect Multimorbidity Patterns

An alternative to the manual grouping of diagnoses is the use of computer algorithms to create groups. A cluster is a group of objects that are similar to each other, whereas objects in different clusters are expected to be far from each other or at least less similar than those in the same cluster [12]. Clustering can be used to detect multimorbidity patterns by grouping either patients or diseases [13]. If we group the diagnoses, one diagnosis belongs to only one cluster, whereas a patient can belong to several clusters. If we group the patients, the reverse is true: one diagnosis can belong to several groups, but one patient can belong to only one cluster. This study focused on grouping diagnoses.

The data used in clustering can be either numerical values or text. Here, we follow the study by Hidalgo et al [14] and represent the diagnoses as nodes and their relationships as links in a network. We refer to this as the *multimorbidity network*. In this network, the weight of the links between 2 diagnoses measures how strongly they correlate in a patient record database.

Although clustering algorithms have been widely used elsewhere in health care, the existing literature lacks reliable, automatic, and computer-generated clusters. Estiri et al [15] used clustering to detect anomalies in health records by combining agglomerative clustering with a k-means algorithm. The idea was to detect small clusters and flag them as anomalies. The authors reported a significantly smaller number of false positive cases than simple anomaly detection based on the SD and Mahalanobis distance.

Huang et al [16] clustered patients into 5 clinically meaningful groups based on the similarity of their diagnoses and the geographical locations of the hospitals. Their motivation was to build machine learning models trained for each group separately to provide a better prediction of mortality and intensive care unit stay time.

Kalgotra et al [17] used co-occurrence statistics to build a *multimorbidity network* to study the disparity of gender. The statistics were extracted from the treatment data of >22.1 million patients. They created networks separately for men and women and compared the structures of the 2 networks. The networks of female patients had more connections with mental health.

Folino et al [18] clustered patients based on a multimorbidity network built using co-occurrence statistics. They used the k-means clustering algorithm with Jaccard distance. A

representative of each cluster was chosen as the set of all diseases whose relative frequency in the cluster exceeded a user-defined threshold (eg, 0.8). Clustering was used to predict future diseases and was tested using the records of 1462 patients from a small town in South Italy.

In the study by Folino and Pizzuti [19], the same prediction system was revised using common neighbors in the network. Records of 2541 patients from 2000 to 2009 were used to build a network from ICD-9-CM codes. The resulting network contained 492 nodes and 21,676 connections. A total of 2 separate subnetworks were created. The first included only connections with *a relative risk* (RR) value of >20 (2330 connections), and the other included those with a Pearson correlation value of ≤0.06 (7242 connections). Future patient diseases were predicted by calculating the number of common neighbors shared by the 2 diseases.

Ding et al [20] extended the previous prediction model using ICD, 10th Revision (ICD-10) and demographic data. On the basis of data collected between 2007 and 2014 in an (unnamed) provincial capital in China, they reported that 71% of acute diseases and 82% of chronic diseases were predictable.

John et al [21] applied clustering to 1039 American Indians using data from an interview-based questionnaire. Cornell et al [20] used ICD-9 codes from data obtained from administrative databases of primary care clinics. Marengoni et al [22] used electronic medical records of the acute care wards of 38 internal medicine and geriatric wards in Italy in 2008.

Marengoni et al [22] calculated clusters of diseases to detect groups of patients at risk of in-hospital death. Their data comprised 1332 older people hospitalized in acute care wards. This small data set had 19 diagnoses, which were grouped into 8 clusters using a correlation matrix and average linkage agglomerative clustering. The results included 4 clusters comprising a disease and its possible consequences. For example, diabetes is clustered with cerebrovascular diseases and coronary heart diseases, thyroid dysfunction with anxiety, and chronic renal failure with anemia. The combination of chronic renal failure and anemia had the highest likelihood of in-hospital death, with an odds ratio of 6.1.

Most existing studies on clustering are based on hierarchical agglomerative methods using heuristic criteria, either *average* or *complete linkage* [13]. Wartelle et al [23] extended hierarchical agglomerative clustering by directly optimizing clustering using RR. By default, this is a more solid approach than any linkage criterion (single, average, or complete). They applied the method to data collected from the emergency department (ED) of Troyes Hospital in Eastern France during a 2-year period between 2017 and 2019. A network comprising 151 ICD-10 blocks was created using 114,391 hospital visits of 72,666 patients.

Proposed Methodology

In this study, instead of agglomerative clustering, we applied a *k-means*–based algorithm. Previously, k-means clustering was used for clustering patients [24]. We applied the algorithm for clustering diseases using data comprising 45 million health care visits covering all public health service use (both primary and secondary care) of the population aged ≥18 years in the entire of Finland from 2015 to 2018. This data set is significantly larger than that used in any of the previous studies.

We constructed a multimorbidity network comprising diseases represented as blocks of the ICD-10 codes. Correlated diseases were in the network. The strength of the links between the diseases was measured using RR, which estimates how much higher the observed prevalence is in relation to the expected prevalence. Clustering was used to find multimorbidity patterns by dividing the network into subgroups with high RR values within. These groups can contain previously unknown multimorbidity patterns.

Similar to the study by Wartelle et al [23], our study was also based on RR. However, there were 2 main differences. First, the agglomerative clustering algorithm in the study by Wartelle et al [23] needs to access the original data after each merge to recalculate the RR values, which is very time consuming with large data. We constructed the network only once, without any need to access the original data after that. This approach scales better as the network is remarkably smaller than the original data (205 nodes vs 58 million patients). K-means itself may require multiple runs [25] to create accurate clustering; however, we avoided this by using a more robust derivation called the M-algorithm [26].

The second difference is that the results of [23] were obtained from emergency visits. Although the resulting clusters could be valid in this context, the generated clusters were different from those obtained from all general health care visits.

The main contributions of our paper can be summarized as follows:

- We use a k-means–based algorithm called M-algorithm, which has been shown to provide highly accurate clustering with controlled validation data sets and scaling up to large-scale data [26].
- We use inverse internal weight (IIW) in the network as a cost function as it has been shown to provide more balanced cluster sizes than other alternatives [26].
- We apply the algorithm to large-scale data comprising 58 million health care visits in all of Finland from 2015 to 2018.
- We make the data publicly available on the University of Eastern Finland website [27], including the multimorbidity network and the clusters.

These contributions directly support several of the goals described by Whitty and Watt [28]. These objectives include strengthening statistical methods to detect clusters, applying them to large data sets, and treating clusters of diseases more effectively. In this paper, we describe the content of the generated clusters and their relationships with nearby clusters. We report the most significant observations and their effects on both service use and costs in the health care system. The study follows the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) guidelines [29] for all relevant items except those related to prediction.

Methods

Overview

Graph clustering has been used in physics [30,31], engineering [32], image processing [33], and medical [34] and social sciences [35]. The technique has several names, including *network community detection* [36-42], *graph clustering* [43] or *graph partitioning* [33,44,45]. These methods can be directly applied to diseases by considering the co-occurrence matrix of diseases as a graph.

By grouping data into meaningful clusters and finding co-occurring diagnoses, it is possible to plan the treatment processes of multimorbid patients and the resources needed in service provision. It is known that diseases often cluster because of a common risk factor; however, only a small number of possible clusters and the connections between the clusters are well known [28].

Data

A summary of the patient record database is presented in Table 1. The data were extracted from the National Administrative Care Register for Health Care, covering all inpatient and outpatient primary and specialized care between 2015 and 2018. Finnish health care registers include data on the patient's age, gender, and the municipality of residence, as well as information concerning the service event, such as the type of contact (visit, phone call, or inpatient admission) and reason for the visit, treatment, and procedures. Reasons for visits were recorded using ICD-10 or International Classification of Primary Care, second edition codes.

Table 1. Summary of the patient database.

Data	Values
Entire database	
All patients, n (%)	4,280,985 (100)
Patients with ICD-10[a] codes	3,987,382 (93.14)
Time range	2015 to 2018
Total visits, n (%)	311,721,962 (100)
Visits with ICD-10 codes	69,306,854 (22.23)
Number of diagnoses per visit, mean	1.6
Total cost of all visits per year (€[b])	9685 million
Included in clustering	
Visits, n (%)	58,391,604 (18.73)
Costs per year (€)	6596 million
Cost of patient per year (€), mean (SD)	2538 (6478)
Patients, n (%)	3,835,531 (89.59)
Patients per year, mean (SD)	2,536,944 (37,494)
Gender, n (%)	
Women	2,062,110 (54)
Men	1,773,419 (46)
Age (years), median	54
Patients aged >70 years, n (%)	943,717 (25)

[a]ICD-10: International Classification of Diseases, 10th Revision.

[b]A currency exchange rate of €1=US $1.09 is applicable.

The entire patient record database contains information on 4.3 million patients aged >18 years. For the cluster analysis, we only included patients with a medical diagnosis (excluding external cause diagnoses), which totaled 3.8 million. The full database included approximately 312 million contacts with health services. The visits were divided into 272,090,337 contacts with primary care services and 39,631,625 contacts with special care services. Primary care contacts included 142,874,297 home visits, 71,658,708 visits to a health center, 26,849,249 phone calls, and 30,708,083 other types of contacts.

For the clustering analysis, from all the visits (311,721,962), we included only those having ICD-10 diagnoses recorded (n=69,306,854 [22.23%]). We excluded all the symptom codes (R00-R99); external causes for injuries, diseases, and deaths (V01-Y92); and health factors and contacts to the service providers (Z00-ZZB), as they do not represent any disease themselves, as well as special diagnosis codes (U00-U99). After filtering these out, the remaining data included 18.73% (58,391,604/311,721,962) of visits.

The costs for each diagnosis were calculated using the computational standard cost [46,47] using patient grouping methods and standard unit costs calculated from national-level cost accounting projects. Hospitalizations and hospital outpatient visits were grouped using the Nordic Diagnosis-Related Groups grouper. The Nordic Diagnosis-Related Groups cost weights for hospitalizations and outpatient visits were based on individual-level cost accounting data from several hospitals and were used in the national price lists by the Finnish Institute for Health and Welfare [48]. The unit cost estimates for each type of primary care contact were obtained from the national standard price list for primary care encounters. The unit cost estimates for social care encounters and community care bed-days were derived from the national price list for the unit costs of health care services in Finland.

The total annual health service cost in Finland during the period 2015 to 2018 was €9685 million for a total of 311 million visits. A currency exchange rate of €1=US $1.09 is applicable. The cost estimation for the data used in the cluster analysis totals to €6596 million per year. The annual cost of each year had an increasing trend between 2015 and 2017 but decreased in 2018: €6579 million (2015), €6626 million (2016), €6723 million (2017), and €6455 million (2018). Some changes may have originated from changes in recording practices. In addition, patients who were hospitalized for longer periods (weeks or months) were not included in the 2018 data if they were not discharged by the end of 2018.

Measuring RR

There are several possibilities for measuring the strength of the relationship between 2 diseases (Table 2). These include ϕ *correlation* (Pearson correlation) [14,34], *co-occurrence correlation* [49], *Jaccard coefficient* [50], *Yule Q* [21,22], *Salton cosine index* [17], and multiple variants of RR [18,19,26]. For a good review, refer to the study by Srinivasan et al [49].

Table 2. Ways of measuring disease connectivity.

Name	Formula[a]	References
Relative risk 1		[14,51]
Relative risk 2		[18]
Relative risk 3		[52]
Co-occurrence correlation		[49]
ϕ-correlation		[14,18,34] (slight variation [52])

[a]N: number of patients; P_x: number of patients with diagnosis x (prevalence); P_{xy}: number of patients with both diagnosis x and y (prevalence); E[xy]: expected frequency of xy; $p(x)=P_x/N$: probability of a random patient having a diagnosis x; $p(xy)=P_{xy}/N$: probability of a random patient having both diagnosis x and y.

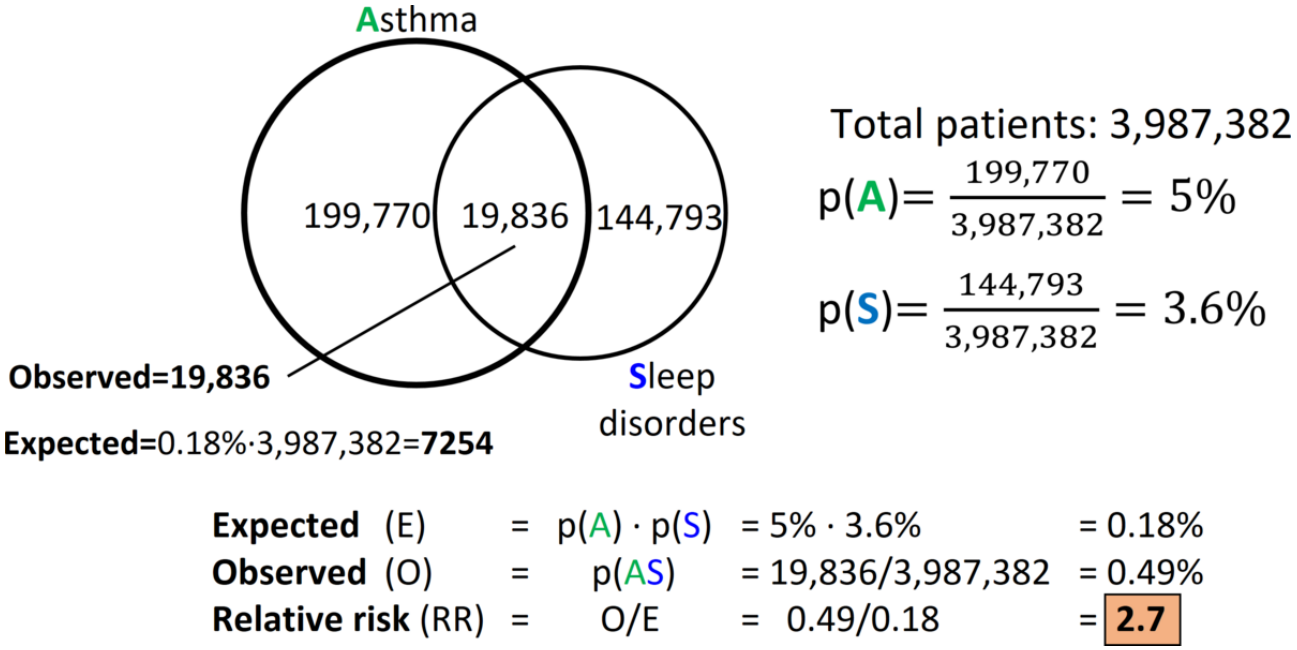
Several authors [17,23,49] have noted that the existing measures contain biases. For example, RR overemphasizes the connection between infrequent diseases. The Pearson correlation underestimates the relationship between common and infrequent diseases. Owing to these problems, Srinivasan et al [49] ended up proposing their own method, called *co-occurrence correlation*.

We used RR (variant 1 in Table 2) as this measure has been widely used in the literature, and its values are clear to understand. It has been used previously by several authors [14,18,23] to study the relationship between diagnoses. It can also be used for other purposes; for example, to study market baskets [51].

RR is defined based on the diagnoses' prevalence, as follows:

Here, $p(x)$ (P_x/N) and $p(y)$ (P_y/N) are the probabilities that a randomly chosen patient has diseases x and y, respectively, and $p(xy)$ (P_{xy}/N) is the probability that a randomly chosen patient has both diseases. $E[xy]$ is the expected frequency of xy. Figure 1 demonstrates the detailed calculation of the RR values in cases of asthma and sleep disorders. An RR value >1.0 indicates that the 2 diseases are related.

Figure 1. Example of measuring comorbidity by relative risk. Here, asthma and sleep disorders are highly correlated. If they were independent of each other, the probability of a person having both should be p(A) × p(B) = 0.18%, whereas their observed co-occurrence would be 0.49%. Therefore, the relative risk to have both is 2.7 times higher than by random chance.

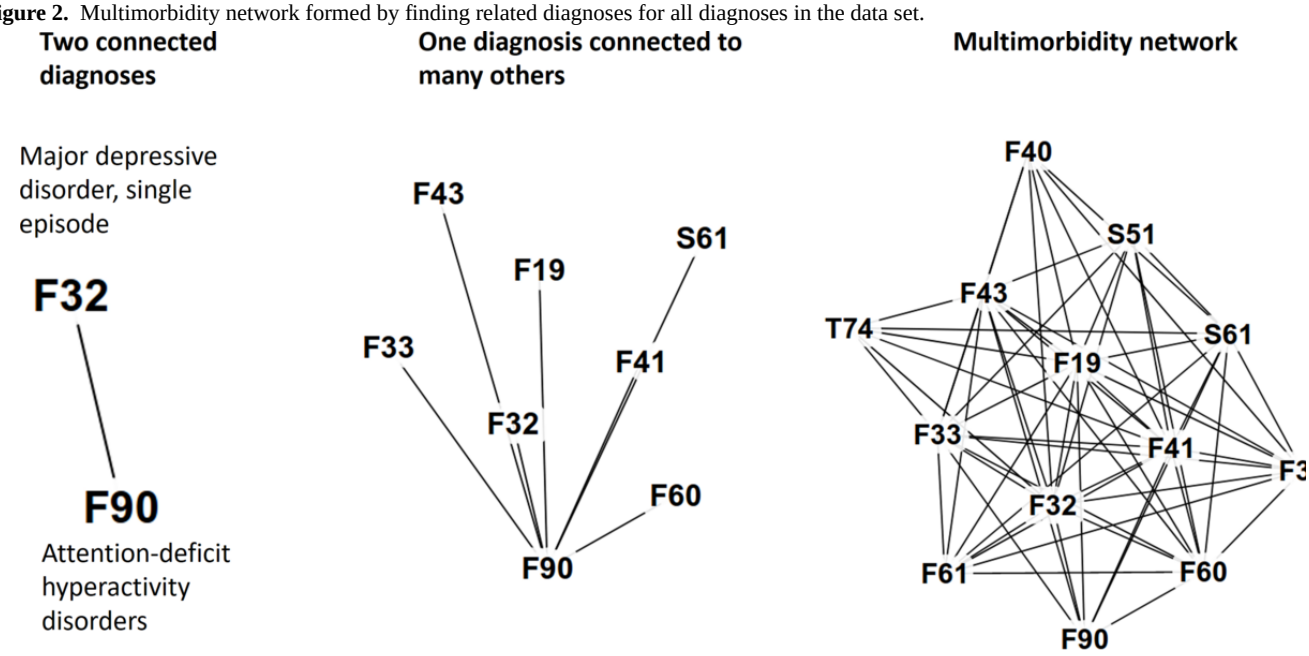


Most RR values are between 0.5 and 5.0; however, they can also be >100. These outlier values would dominate the clustering cost function optimization, and for this reason, we normalized them to the range of (0,1) by using the following variant of the generalized symmetrical sigmoid function [53]:

Multimorbidity Network

A multimorbidity network is formed by connecting all pairs of diagnoses that are related (Figure 2). Each node in this network corresponds to a medical diagnosis, and the strength of the connections can be measured using RR, correlation, or other methods. We used the name multimorbidity network following the choice of Aguado et al [54]. This network has also been called a disease co-occurrence network [48], *phenotypic disease network* [14], *comorbidity network* [17], and *disease comorbidities network* [34].

Figure 2. Multimorbidity network formed by finding related diagnoses for all diagnoses in the data set.

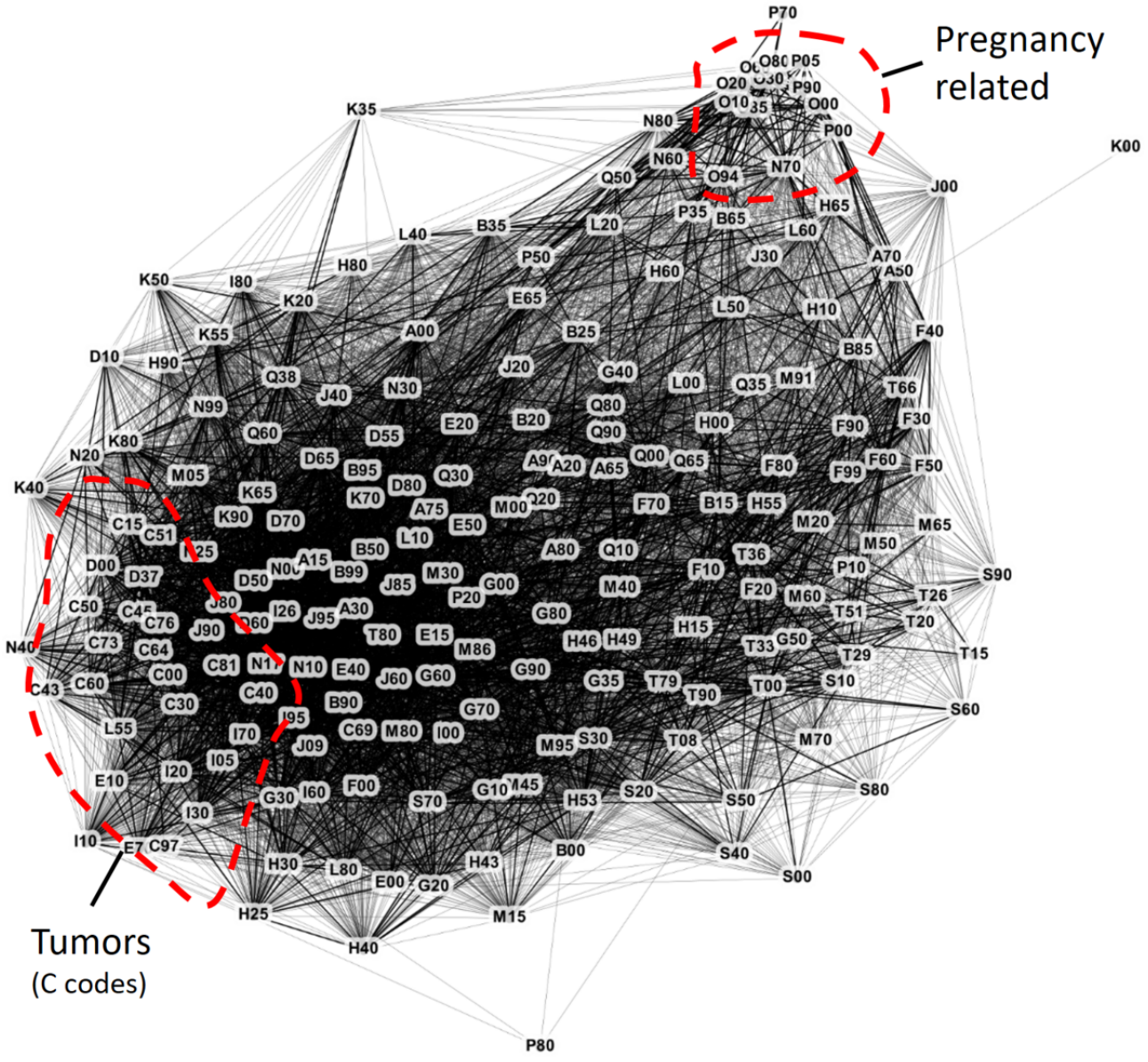


Several previous studies used multimorbidity networks [14,17,18,34,49,54]. In addition, Klimek et al [55] and Moni and Liò [52] studied comorbidity associations, although they did not explore much of the network analysis. Moni and Liò [52] created R language software called *comoR* for disease comorbidity risk analysis. Divo et al [34] studied chronic

obstructive pulmonary disease for disease screening and management. Folino et al [18] predicted future diseases based on past medical history. Srinivasan et al [49] used a multimorbidity network to extract features for a high-cost patient prediction. Hidalgo et al [14] also published multimorbidity network data (based on 13 million patients) [56].

We constructed a multimorbidity network (Figure 3 [57]) by calculating the RR value for all pairs of diagnoses, including those with an RR value ≥1.0 and at least 10 patients with both diagnoses. The accuracy used for diagnoses was the subgroup of the ICD-10 classification (eg, I20-I25). We also filtered out the diagnoses that indicated symptoms and external causes (those starting with Z, W, Y, and R). After filtering, we obtained 205 disease subgroups in the graph (see Multimedia Appendix 1 for the full list).

Figure 3. The full network was overwhelming to analyze, with 205 disease subgroups and 14,254 connections overall. Here, we show only the 8895 connections with a relative risk of >1.5. Connections with relative risk >3.0 are drawn in bold. ICD-10 (International Classification of Diseases, 10th revision) subgroups are represented by the first diagnosis of the group (Multimedia Appendix 1). The image was created by using the Gephi software [56]. Only very tight groups such as pregnancy-related diagnoses and tumors can be recognized from the network.



Clustering

Overview

The main motivation for clustering is that the multimorbidity network is too large (205 nodes and 14,254 connections) for detailed analysis. For this reason, we clustered the graph to form more compact entities of related diseases. The goal was to assign strongly related diseases to the same cluster but keep uncorrelated diseases in different clusters. To achieve this goal, an evaluation criterion was necessary to measure the effectiveness of clustering.

Cost Function

Instead of using heuristic criteria such as average or complete linkage, it is better to define an exact cost function that the clustering algorithm optimizes directly. When clustering numerical data, a typical goal is to measure the compactness of

the clusters. For example, both the Ward method and k-means minimize the *sum of squared distances* between the data objects to the cluster mean (*centroid*). However, calculating the mean of a subgraph is not possible directly but would require an indirect solution such as vectorizing the nodes by graph embedding [58]. Moreover, calculating the distance between 2 nodes is not possible if they are not connected. Therefore, graph-specific cost functions have been developed to overcome these issues.

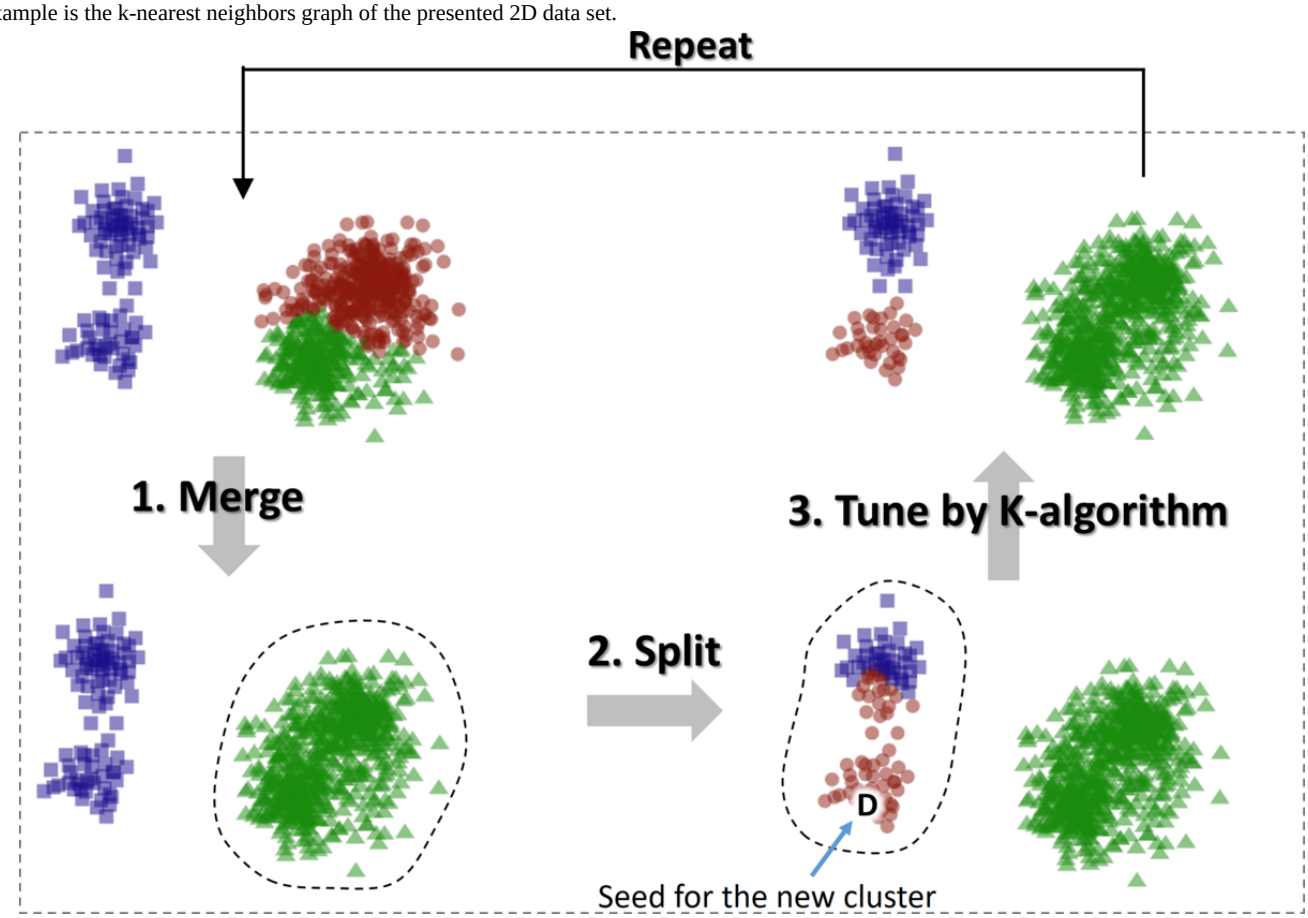
Three cost functions were evaluated in the study by Sieranoja and Fränti [26] with controlled data—*conductance*, *mean internal weight*, and *IIW*. The last function produced the most accurate clustering result with balanced cluster sizes and was therefore chosen in this study as well. When k is the number of clusters, W_i is the internal weight of cluster i, and M is the total weight (mass) of the entire graph, the cost is calculated as follows:

In multimorbidity network analysis, it is desirable to have clusters of approximately the same size. This could be controlled by specifying the number of clusters. As the cost function induces balanced cluster sizes, we aimed to group N nodes into k clusters of size N/k=n. In our case, we had N=205 diseases and k=15 clusters with 205/15=13.7 diseases, on average. This size was sufficiently small to allow us to investigate the clusters manually.

Clustering Algorithm

We used the recently developed M-*algorithm* in [26], which combines a k-means type of iterative optimization with an additional merge and split strategy to escape from local minima (Figures 4-5). The *IIW* was the recommended cost function.

Figure 4. The M-algorithm merges 2 random clusters, splits 1 random cluster, and fine-tunes the result by using the K-algorithm. The network in this example is the k-nearest neighbors graph of the presented 2D data set.



K-means uses two optimization steps: assignment and centroid steps. In the assignment step, every point is placed in the cluster whose mean (centroid) is closest. However, the assignment of points is not independent of the assignment of other points. Their joint effect may cause the cost value to fluctuate so that the total value increases even if the single assignment decreases. To avoid this problem, we used the sequential variant of k-means, where every assignment has an immediate effect on the centroids. This technique prevents fluctuations.

The k-means variant applied to graphs is called the *K-algorithm*, which is similar to the original k-means algorithm but without centroids. The distance calculations were replaced by directly evaluating the effect of the assignment on the cost function. Most cost functions are based on maximizing the weights inside the cluster or minimizing external weights. Therefore, the effect of a node joining a cluster can be calculated using only its edges and the size of the cluster.

The K-algorithm iteratively improves the initial solution by sequentially processing the nodes in random order. For each

node, the method considers all clusters and checks whether changing the partition of the node improves the cost function. If it does, the cluster assignment is changed. After all the nodes have been processed, the algorithm starts another iteration. The iterations continue until no changes occur.

The M-algorithm differs from the K-algorithm in the additional merge and split step. The M-algorithm first merges 2 random clusters and then splits 1 random cluster. The clustering solution is fine-tuned using the K-algorithm. If the new solution improves the cost function value, it is kept as the current solution; otherwise, the process continues from the previous solution. The merge and split process is repeated depending on the amount of computation time required. The pseudocode for the algorithm is presented in Figure 5.

Figure 5. Pseudocode for the M-algorithm.

```
MergeAndSplit(graph,k,R)
INPUT:
   graph (with N nodes)
   k = number of clusters
   R = number of repeats
 1 cluster = K_algo(graph,NULL,k)
 2 FOR i=1:R
 3   newClu = cluster
 4   (A,B) = choose randomly
 5   newClu = MERGE(newClu, A,B)
 6
 7   // Perform a split
 8   cluId = RAND(1,k)
 9   unbalanceFactor = RAND(0.05,0.95)
10   growSize = unbalanceFactor*SIZE(cluId)
11   seedId = random node from cluster cluid
12   newClu = GrowCluster(graph,newClu,cluId,seedId,growSize)
13
14   newClu = K_algo(graph,newClu,k)
15   IF cost(graph,newClu) > cost(graph,cluster) // improvement
16      cluster = newClu
17 RETURN cluster
```

Merge clusters A and B

Split cluster

Tune by K-algorithm

As the network itself is quite small (205 diagnoses), the clustering algorithm takes only a little time. The time complexity of the M-algorithm is $O(RIN[k+|E|/N])$, where R is the number of repeats, N is the number of diagnoses (nodes), k is the number of clusters, $|E|/N$ is the average number of connections for each node (diagnosis), and I is a small number that reflects the number of iterations to converge. We ran the M-algorithm for 20,000 repeats, which took 27 minutes (single thread) on an Intel Xeon(R) W-2255 CPU at 3.70 GHz. The bottleneck was the $O(N_v)$ network construction, which needed to process all N_v=58 million patient visits and took 52 minutes.

The number of clusters, k, must be fixed by the researcher beforehand. A small number is likely to generate large mixed clusters of many diseases, thereby losing the capability to make meaningful observations. A large number of clusters tend to mainly cluster diseases from the same ICD group, which might lose the chance to detect relevant multimorbidity patterns. We tried clustering with several different k values and chose k=15 as it produced clusters of convenient size for analysis in the form of similarity matrices.

It is also possible for the algorithm to recommend the number of clusters using a suitable cluster validity index that measures the ratio of within-cluster and between-clusters similarities, as in the study by Zhao and Fränti [59]. Wartelle et al [23] derived a validity index from RR and obtained k=16 clusters in their data. We used the *silhouette coefficient* [60] for our data, and in the range of 5 to 25, it obtained k=17 clusters. They are both close to our choice of k=15.

Ethics Approval

Permission to use the register data was obtained from the Finnish Institute for Health and Welfare. All methods were carried out in accordance with relevant guidelines and regulations or declaration of Helsinki. The Finnish legislation (Act 552/2019) do not require informed consent for register-based research when study is solely based on registers and the study is considered to be of public health importance.

Results

RR Measurements

Table 3 shows the 10 pairs of disease subgroups with the highest RR values. They are diagnoses with the highest probability of appearing jointly relative to the expected probability with the independent assumption. Some connections are obvious, often

representing the same or closely related conditions (C40-C41 and C45-C49). Some have known explanations in medical science (F70-F79 and Q90-Q99) or a clear causal relationship (D80-D89 and N00-N08). There are also connections with smaller RR values that are not so obvious at first sight; however, they are clinically meaningful (I26-I28 and M30-M36). In addition to using the ICD-10 subgroups, we calculated the RR values for diagnoses with 3-character precision. Some RR values <1.0 were also found for diagnoses such as E10 and E11, which are exclusive to each other.

Table 3. The 10 disease pairs with the highest relative risk (RR) valuea.

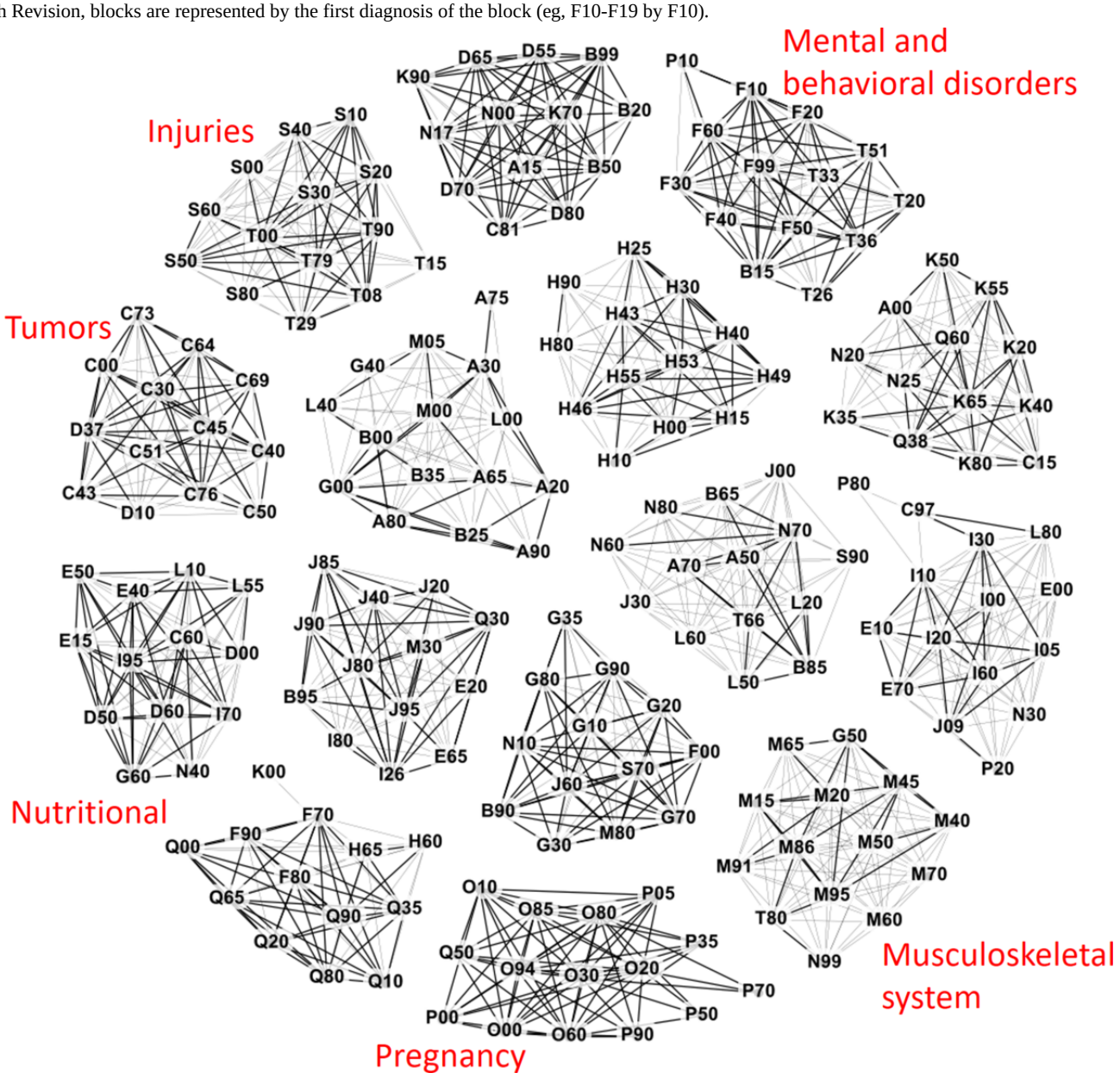
Diagnosis A		Diagnosis B		RR	Count (n=3987, 382%), %
Code	Description	Code	Description		
A80-A89	Viral infections of the central nervous system	G00-G09	Inflammatory diseases of the central nervous system	170.1	484 (0.01)
A15-A19	Tuberculosis	B90-B94	Sequelae of infectious and parasitic diseases	110.7	132 (0.00)
C40-C41	Malignant neoplasms of bone and articular cartilage	C45-C49	Malignant neoplasms of mesothelial and soft tissue	98.3	107 (0.00)
T20-T25	Burns and corrosions of external body surface, specified by site	T29-T32	Burns and corrosions of multiple and unspecified body regions	91.0	893 (0.02)
F70-F79	Mental retardation	Q90-Q99	Chromosomal abnormalities, not elsewhere classified	79.7	945 (0.02)
G35-G37	Demyelinating diseases of the central nervous system	H46-H48	Disorders of optic nerve and visual pathways	50.7	811 (0.02)
D80-D89	Certain disorders involving the immune mechanism	N00-N08	Glomerular diseases	47.2	2386 (0.06)
J85-J86	Suppurative and necrotic conditions of lower respiratory tract	J90-J94	Other diseases of pleura	45.7	866 (0.02)
N25-N29	Other disorders of kidney and ureter	Q60-Q64	Congenital malformations of the urinary system	45.3	328 (0.01)
F70-F79	Mental retardation	Q00-Q07	Congenital malformations of the nervous system	42.0	238 (0.01)

[a]Full list is available on the University of Eastern Finland website [27].

Clustering Results

The overall clustering results are visualized as a graph in Figure 6. The graph shows connections within the clusters; however, all connections between clusters have been eliminated for clarity.

Figure 6. Clusters obtained from the multimorbidity network. Subjective labels of 6 clusters are also shown. This figure shows all 205 diagnoses and only those 1144 connections with relative risk ≥1.5. Cases with a relative risk of ≥3 are shown with thicker lines. International Classification of Diseases, 10th Revision, blocks are represented by the first diagnosis of the block (eg, F10-F19 by F10).



We fixed the number of clusters to 15 for the M-algorithm [26]. This roughly matches the number 16 used in a study by Wartelle et al [23]. The main characteristics of the resulting clusters are summarized in Tables 4 and 5. The strength of the associations between the diagnosis subgroups inside the 2 example clusters and the connections between the 2 clusters can be observed in Figure 7. The number of patients in each cluster, the number of visits to health services, total costs, cost per visit, and cost per patient are reported in Table 6.

Table 4. Content of the 15 clusters (ICD-10a blocks) and their strengths as the mean RRb values of diagnoses within the cluster.

Cluster	RR, mean	ICD-10 codes
Cluster 1	11.3	O85-O92; O30-O48; O20-O29; O10-O16; O60-O75; O94-O99; O80-O84; P05-P08; P00-P04; O00-O08; P35-P39; P90-P96; Q50-Q56; P70-P74; P50-P61
Cluster 2	8.1	B50-B64; N00-N08; D70-D77; C81-C96; D55-D59; D80-D89; D65-D69; B99-B99; A15-A19; N17-N19; B20-B24; K70-K77; K90-K93
Cluster 3	7.8	F70-F79; Q90-Q99; F80-F89; Q00-Q07; Q35-Q37; Q80-Q89; Q65-Q79; F90-F98; Q20-Q28; Q10-Q18; H65-H75; H60-H62; K00-K14
Cluster 4	7.6	C40-C41; C45-C49; C76-C80; C30-C39; D37-D48; C69-C72; C00-C14; C51-C58; C64-C68; C73-C75; C50-C50; C43-C44; D10-D36
Cluster 5	5.7	J95-J99; J85-J86; J90-J94; J80-J84; Q30-Q34; I26-I28; J40-J47; B95-B98; M30-M36; J20-J22; E65-E68; E20-E35; I80-I89
Cluster 6	5.4	T36-T50; B15-B19; F60-F69; F10-F19; F99-F99; T51-T65; F20-F29; F30-F39; T33-T35; T26-T28; F40-F48; T20-T25; F50-F59; P10-P15
Cluster 7	4.8	E40-E46; E50-E64; D60-D64; I95-I99; D50-D53; L55-L59; D00-D09; E15-E16; G60-G64; I70-I79; L10-L14; C60-C63; N40-N51
Cluster 8	4.6	G80-G83; G10-G14; J60-J70; G90-G99; F00-F09; G70-G73; G30-G32; N10-N16; B90-B94; S70-S79; M80-M85; G20-G26VG35-G37
Cluster 9	4.5	Q60-Q64; N25-N29; K65-K67; Q38-Q45; C15-C26; K80-K87; K55-K64; K40-K46; N20-N23; K20-K31; K50-K52; A00-A09; K35-K38
Cluster 10	4.3	G00-G09; A80-A89; A90-A99; A65-A69; M00-M03; A30-A49; B25-B34; A20-A28; M05-M14; B00-B09; L00-L08; L40-L45; B35-B49; G40-G47; A75-A79
Cluster 11	3.8	H53-H54; H46-H48; H55-H59; H49-H52; H43-H45; H30-H36; H15-H22; H40-H42; H25-H28; H00-H06; H10-H13; H90-H95; H80-H83
Cluster 12	3.0	T00-T07; T90-T98; T79-T79; S10-S19; S30-S39; S20-S29; T08-T14; T29-T32; S50-S59; S40-S49; S80-S89; S60-S69; S00-S09; T15-T19
Cluster 13	2.9	M95-M99; M40-M43; M45-M49; M86-M90; T80-T88; G50-G59; M15-M19; M20-M25; M50-M54; M91-M94; M65-M68; M70-M79; N99-N99; M60-M63
Cluster 14	2.9	A50-A64; A70-A74; B85-B89; N70-N77; B65-B83; T66-T78; L50-L54; L20-L30; N80-N98; L60-L75; J30-J39; N60-N64; J00-J06; S90-S99
Cluster 15	2.1	I30-I52; I20-I25; I60-I69; I10-I15; L80-L99; I05-I09; J09-J18; E70-E90; N30-N39; E10-E14; E00-E07; I00-I02; P20-P29; C97-C97; P80-P83

[a]ICD-10: International Classification of Diseases, 10th Revision.

[b]RR: relative risk.

Table 5. Summarization of the cluster content with their age and gender distributions.

Cluster	Dominant gender		Age (years), median	Age ≥70 years, %	Description
	Gender	Values, n (%)			
Cluster 1: pregnancy	Women	219,566 (99.68)	33	0	Pregnancy, childbirth and the puerperium (O codes), certain conditions and disorders originating in perinatal period (P05-P08, P00-P04, P35-P39, P90-P96, P70-P74, and P50-P61), and congenital malformations of genital organs (Q50-56)
Cluster 2: immune system and blood-forming organs	Men	110,157 (50.79)	69	50	Infectious diseases strongly affecting the immune system (B50-B64, B20-24, B99-B99, and A15-19); malignant neoplasms of lymphoid, hematopoietic, and related tissue (C81-96); diseases of the kidneys (N00-N08 and N17-N19), liver (K70-77), blood, and blood-forming organs and disorders of the immune mechanism (D70-D77, D55-D59, D80-D89, and D65-D69 [except nutritional and aplastic and other anemias]); and other diseases of the digestive system (K90-K93)
Cluster 3: mixed cluster; includes mental disorders, malformations, and ear and oral cavity diseases	Women	1,062,480 (55.13)	49	17	Mental retardation (F70-79) and disorders of psychological development or unspecified disorder (F80-F89, F99-F99) and congenital malformations (Q codes except for codes for congenital malformations of the respiratory system, digestive system, genital organs, and urinary system); diseases of the ear (H65-H75 and H60-H62); and diseases of the oral cavity, salivary glands, and jaws (K00-K14)
Cluster 4: tumors	Women	317,372 (62.64)	66	42	Malignant neoplasms (all C codes, except codes for malignant neoplasms in digestive organs; male genital organs; lymphoid, hematopoietic, and related tissue; multiple independent sites) and benign neoplasms (D10-D36)
Cluster 5: lower respiratory system	Women	437,591 (59.13)	64	38	Lower respiratory tract diseases and related inflammatory conditions (J95-J99, J85-J86, J90-J94, J80-J84, J40-J47, and J20-J22); congenital malformations of the respiratory system (Q30-Q36), pulmonary heart disease and diseases of pulmonary circulation (I26-I28); bacterial, viral, and other infectious agents (B95-B98); systemic connective tissue disorders (M30-M36), obesity (E65-E68) and disorders of other endocrine glands (E20-E35); and diseases of veins, lymphatic vessels, and lymph nodes not classified elsewhere (I80-I89)
Cluster 6: mental and behavioral disorders	Women	369,203 (58.30)	46	15	Mental and behavioral disorders and substance abuse problems (F60-F69, F10-F19, F20-F29, F30-F39, F40-F48, F50-F59, and F99); poisonings (T36-T50 and T51-T65) and certain viral infections (B15-B19); and related burns (T20-T25 and T26-T28), frostbite injuries (T33-T35), and birth trauma (P10-P15)
Cluster 7: nutritional	Men	314,390 (66.98)	72	58	Malnutrition (E40-E46) and nutritional deficiencies (E50-64); anemias (D50-D53 and D60-D64); other and unspecified disorders of the circulatory system (I95-I99); certain skin diseases (L55-L59 and L10-L14); in situ neoplasms (D00-D09); other disorders of glucose regulation and pancreatic internal secretion (E15-E16); polyneuropathies (G60-G64); diseases of arteries, arterioles, and capillaries (I70-I79); and diseases and malignant neoplasms of male genital organs (C60-C63 and N40-N51)
Cluster 8: diseases related to aging	Women	242,917 (59.83)	76	64	Cerebral palsy, memory disorders, other diseases of the central nervous system or neurodegenerative diseases (included G-codes), lung diseases because of external agents (J60-J70), organic mental disorders (F00-F09), renal tubulointerstitial diseases (N10-N16), changes in bone structure (M80-85) and injuries (hip and thigh S70-S79), and other infections (B90-B94)

Cluster	Dominant gender		Age (years), median	Age ≥70 years, %	Description
	Gender	Values, n (%)			
Cluster 9: mixed cluster; includes organ malformations and digestive system disorders	Women	387,222 (54.07)	63	37	Congenital malformations of the urinary system and digestive system (Q60-Q64 and Q38-Q45), some disorders of the kidney and ureter (N25-N29) and genitourinary system (N20-N23), diseases of the digestive system (all K codes, except diseases of the oral cavity, salivary glands and jaw, and diseases of the liver), malignant neoplasms of digestive organs (C15-C26), and intestinal infectious diseases (A00-A09)
Cluster 10: infections and inflammation	Women	483,595 (53.55)	61	33	Inflammatory diseases (G00-G09)/viral infections (A80-A89) of the central nervous system, hemorrhagic fevers (A90-A99), certain other infectious and parasitic diseases (A65-A69, A30-A49, A20-A28, A75-A79, B00-B09, and B35-B49), infectious arthropathies/inflammatory polyarthropathies (M00-M03 and M05-M14), infections of the skin and subcutaneous tissue or papulosquamous disorders (L00-L08 and L40-L45), and episodic and paroxysmal disorders (G40-G47)
Cluster 11: eye and ear	Women	491,892 (58.89)	67	45	Diseases of the eye and adnexa (all H codes) and diseases of the inner ear (H80-H83) and other disorders of the ear (H90-H90)
Cluster 12: injuries	Men	516,849 (51.27)	55	26	Injuries in different parts of the body (all S codes, except injuries to the hip and thigh) and in multiple body regions (T00-T07) or unspecified parts (T08-T14 and T29-T32), effects of foreign bodies entering through a natural orifice (T15-T19), and some of their consequences (T79-T79 and T90-T98)
Cluster 13: musculoskeletal system	Women	855,218 (58.69)	60	31	Diseases of the musculoskeletal system and connective tissue (all M codes, except infectious and inflammatory arthropathies or poly arthropathies, systemic connective tissue disorders, and disorders of bone density and structure); complications of surgical and medical care (T80-T88); nerve, nerve root, and plexus disorders (G50-G59); and other disorders of the genitourinary system (N99-N99)
Cluster 14: mixed cluster; includes sexually transmitted, parasitic, and urinary tract diseases	Women	844,339 (65.79)	48	19	Sexually transmitted diseases (A50-A64 and A70-A74), parasitic diseases (B85-B89 and B65-B83), unspecified effects of external causes (T66-T78), inflammatory diseases of female pelvic organs (N70-N77), disorders of the breast (N60-N64), noninflammatory disorders of the female genital tract (N80-N98), some diseases of the skin (L50-L54, L20-L30, and L60-L75), acute and some other upper respiratory infections (J30-J39 and J00-J06), and injuries to the ankle and foot (S90-S99)
Cluster 15: cardiovascular and metabolic	Women	867,133 (56.22)	68	47	Diseases of the circulatory system (all I codes, except pulmonary heart disease and diseases of pulmonary circulation [I26-I28] and diseases of arteries and veins [I70-I79, I 80-I89, and I95-I99]), other disorders of the skin and subcutaneous tissue (L80-L99), influenza and pneumonia (J09-J18), metabolic disorders (E70-E90), disorders of the thyroid gland (E00-E07), diabetes mellitus (E10-E14), other diseases of urinary system (N30-N39), respiratory and cardiovascular disorders specific to the perinatal period (P20-P29), malignant neoplasms of independent (primary) multiple sites (C97-C97), and conditions involving the integument and temperature regulation of fetus and newborn (P80-P83)

Figure 7. Two example clusters and their connections in between. The numbers are relative risk values. High values and the red color signify stronger relationships. The blocks are represented by the first diagnosis code (eg, T36 represents block T36-T50).

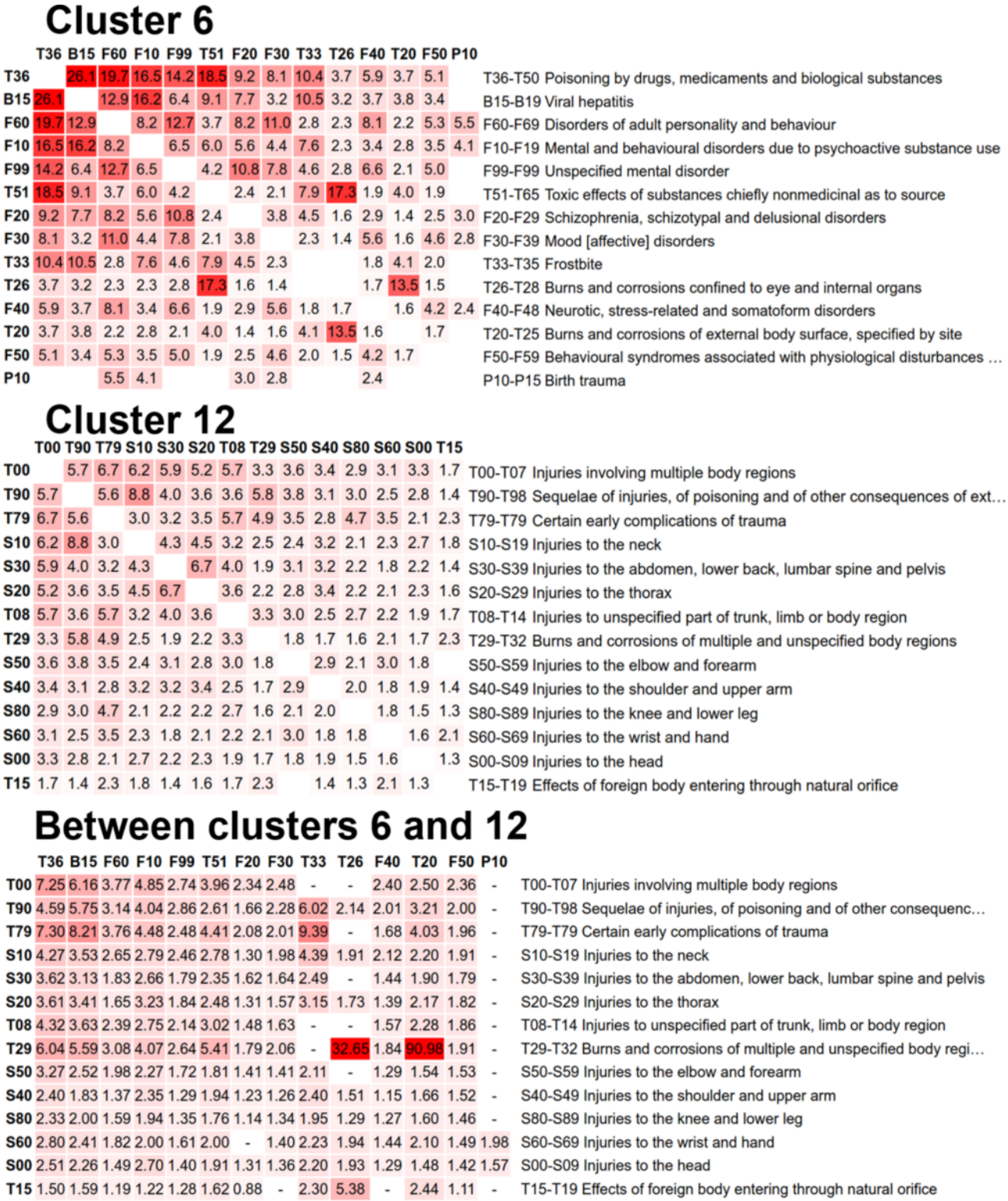


Table 6. Estimated (annual) costs of each clustera.

Cluster	Description	Patients[b] (n=2,536,944), n (%)	Visits[b] (n=14,597,901), n (%)	Total cost[b] (€[c]; millions)	Cost per visit[b] (€)	Cost per patient[b] (€)
1	Pregnancy	78,159 (3.08)	255,902 (1.75)	207	810	2648
2	Immune system and blood-forming organs	95,865 (3.78)	653,500 (4.48)	521	798	5435
3	Mental disorders, malformations, ear and mouth	838,208 (33.04)	1,899,209 (13.01)	324	171	387
4	Tumors	210,272 (8.29)	1,046,147 (7.17)	704	673	3348
5	Lower respiratory system	299,482 (11.80)	953,199 (6.53)	620	651	2070
6	Mental and behavioral disorders	280,450 (11.05)	2,094,496 (14.35)	908	434	3238
7	Nutritional	194,250 (7.66)	708,930 (4.86)	525	741	2703
8	Diseases related to aging	172,194 (6.79)	1,105,325 (7.57)	730	661	4239
9	Organ malformations and digestive system	262,362 (10.34)	867,971 (5.95)	720	829	2744
10	Infections and inflammation	359,738 (14.18)	1,110,728 (7.61)	627	564	1743
11	Eye and ear	320,947 (12.65)	827,680 (5.67)	298	359	929
12	Injuries	324,191 (12.78)	720,282 (4.93)	417	579	1286
13	Musculoskeletal system	616,550 (24.30)	1,704,486 (11.68)	836	490	1356
14	Sexually transmitted, parasitic, urinary tract	474,604 (18.71)	955,465 (6.55)	290	303	611
15	Cardiovascular and metabolic	773,406 (30.49)	3,326,018 (22.78)	2258	679	2920

[a]A patient and a visit can belong to multiple clusters. Visits and costs include only visits and related costs for diagnoses in a cluster. The cost per visit is calculated as an average for the whole 4-year period; all other values are annual.

[b]Number of patients: mean 353,378; number of visits: mean 1,215,289; cost: mean €666 million; cost per visit: mean €583; cost per patient: mean €2377.

[c]A currency exchange rate of €1=US $1.09 is applicable.

Most clusters were dominated by records of female patients. Cluster 1 (219,566/220,280, 100%) included only women, as it comprised pregnancy-related diagnoses. Other clusters with >60% of records of women were cluster 14 (844,339/1,283,478, 65.7%) of mixed diseases (sexual and urinary) and cluster 4 (317,372/506,660, 62.6%) of malignant tumors. The only cluster with a significantly higher proportion of diagnoses from men was cluster 7 (314,390/469,378, 66.9%), which comprised diagnoses mainly related to nutrition. In most other clusters, the proportions of men and women were approximately equal.

The main reasons for female dominance were that the full database included 1,999,325 men and 2,253,669 women and that women had an average of 6.6 diagnoses, whereas men had only 5.4 diagnoses. A possible reason is that there is a lower threshold for women to seek help from health services than for men. For example, the study by Corrigan [61] suggested that social factors discourage men from seeking mental health care, which can lead to the absence of mental health–related multimorbidities among men.

As all diagnoses were forced to belong to a cluster, there were several mixed clusters. For example, the largest cluster (cluster 3) comprised 33.04% (838,208/2,536,944) of patients, including those with dental health problems (K00-K14). If this subgroup of diagnoses were removed, the number of patients would decrease to only 87,634 and would mainly comprise diagnoses related to mental retardation, congenital malformations, and chromosomal abnormalities. However, it is quite logical that dental health–related diagnoses are clustered with mental retardation; congenital malformations; and abnormalities, such as patients with malformations in the oral cavity, jaws, and teeth, which is a patient group treated in the public health service system.

The second-largest cluster (cluster 15), comprising 30.49% (773,406/2,536,944) of patients, included cardiovascular, endocrine, and metabolic diseases. It also had the highest number of visits to health care (3.3 million annual visits). The third-largest cluster (cluster 13) had 24.30% (616,550/2,536,944) of patients but was more focused on diagnoses related to diseases of the musculoskeletal system and connective tissues. Other more clearly focused clusters included tumors (cluster 4), mental disorders (cluster 6), injuries (cluster 12), diseases related to nutrition (cluster 7), and pregnancy (cluster 1). These clusters can be easily explained based on morbidity and mortality data in Finland. Cardiovascular diseases are still the major cause of death [62], and mental disorders are the main cause of disability pensions, followed by musculoskeletal disorders [63].

These clusters also had clear age profiles. The average age of most clusters was rather high, being ≥60 years in the case of 10 clusters. The exceptions were cluster 6 (mental; mean 46 years), cluster 12 (injuries; mean 55 years), mixed clusters 3 (mental, ear, and oral cavity; mean 49 years) and 14 (sexual and urinary; mean 48 years), and cluster 1 (pregnancy; mean 33 years).

Although clustering captures many connections between diseases, it does not capture all information. In fact, many interesting connections can be found by analyzing how strongly the clusters are connected to each other (Figure 8). Cluster 7 (nutritional problems) was the most central cluster, with a strong connection to 10 other clusters. Cluster 1 (pregnancy) was also connected to cluster 6 (mental and behavioral disorders). For example, pregnancy with abortive outcomes (O00-O08) had 5 connections with RR >2 to cluster 6 (mental and behavioral disorders), including neurotic, stress related, mood disorders, and drug poisoning (T36-T50).

Figure 8. Connections between clusters. Each cluster is represented in the rows with the number and description and in the columns with the number. Values in the table represent the number of links with a relative risk of >2.0 between the clusters. Higher values signify a stronger connection and are emphasized by the red color. Three clusters with the highest values for each row are highlighted with bold font.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Pregnancy	-	16	14	3	9	**21**	0	1	13	6	1	1	4	**63**	10
2 Immune system and blood-forming organs	16	-	9	86	**122**	20	**129**	97	**100**	89	41	31	40	13	83
3 Mental disorders, malformations, ear and mouth	14	9	-	10	23	**35**	9	**30**	15	25	**32**	8	22	22	10
4 Tumors	3	**86**	10	-	**75**	3	**98**	63	54	26	23	2	31	6	68
5 Lower respiratory system	9	**122**	23	75	-	15	**115**	93	85	71	34	36	62	11	**96**
6 Mental and behavioral disorders	21	20	35	3	15	-	30	**36**	6	22	7	**89**	16	**54**	4
7 Nutritional	0	**129**	9	98	115	30	-	**132**	98	64	64	56	66	14	**123**
8 Diseases related to aging	1	**97**	30	63	93	36	**132**	-	58	53	46	72	57	1	**100**
9 Organ malformations and digestive system	13	**100**	15	54	**85**	6	**98**	58	-	28	14	8	39	9	58
10 Infections and inflammation	6	**89**	25	26	**71**	22	**64**	53	28	-	26	21	41	38	44
11 Eye and ear	1	41	32	23	34	7	**64**	**46**	14	26	-	7	30	19	**45**
12 Injuries	1	31	8	2	36	**89**	**56**	**72**	8	21	7	-	48	27	18
13 Musculoskeletal system	4	40	22	31	**62**	16	**66**	**57**	39	41	30	48	-	14	46
14 Sexually transmitted, parasitic, urinary track	**63**	13	22	6	11	**54**	14	1	9	**38**	19	27	14	-	8
15 Cardiovascular and metabolic	10	83	10	68	**96**	4	**123**	**100**	58	44	45	18	46	8	-
Sum:	162	876	264	548	847	358	**998**	839	585	554	389	424	516	299	713

Cluster 12 (injuries) had strong connections with clusters 6, 7, and 8. For example, the connection to the nutritional problems cluster had 56 links, with an RR >2. Of these links, 9 came from connections to other and unspecified disorders of the circulatory system (I95-I99).

Figure 7 shows the connections between clusters 6 and 12 in more detail. Cluster 6 comprised mental health (eg, F30-F39 and F60-F69) and substance abuse–related (T36-T50 and F10-F19) diagnoses. Cluster 12 comprised fractures and other injuries. These clusters had a strong connection. A possible explanation is that mental health and substance abuse problems often lead to painful, fracture-causing accidents.

Cost Effect

The costs of all visits, ward stays, and other contacts of patients belonging to the cluster were calculated for those contacts in services with a diagnosis belonging to the cluster. The estimated costs for each cluster are presented in Table 5. The costs are in euro currency (€).

In general, the cost depends on the number of patients and visits. The largest cluster (cardiovascular and metabolic cluster 15) had 3.3 million visits and €2.3 billion in total costs. However, the cost per patient (€2920) was not the highest, and the cost per visit (€679) was only slightly above average. The diseases in the cluster, such as cardiovascular and metabolic disorders, are largely treated in primary health care, and thus, the average visit cost remains relatively low.

For each patient, the highest costs were in cluster 2 (€5435), including infectious diseases strongly affecting the immune system, diseases of the blood and blood-forming organs, and other disorders involving the immune mechanism. These diseases are likely to need frequent contact with specialized care. Per-patient costs were also high in cluster 8 (diseases related to aging), including diagnoses of neurodegenerative diseases and memory disorders requiring frequent health care contacts and intensive care. The cheapest clusters per patient were cluster 3 (mental disorders, malformations, and ear and mouth; €387) and cluster 14 (sexually transmitted, parasitic, and urinary tract diseases; €611). However, if dental diagnoses were removed, the cost for cluster 3 would be €1144.

The highest cost per visit (€829) was in cluster 9, including organ malformations and diseases of the digestive system. The second-highest cost per visit was observed in cluster 1 (pregnancy), where the cost per visit was €810. This is likely because of delivery-related hospital stays, operations, and other specialized care. Regular maternity care visits are not usually recorded using the ICD-10 codes. Clusters with the lowest cost per visit were the same as those with the lowest cost per patient.

Table 7 shows how the costs of some clusters have developed during the years relative to the total cost of all clusters in the same year. Only clusters with a visible trend (increasing or decreasing) are shown. Clusters that included tumors, lower respiratory system, and eye and ear steadily increased their

proportion of all costs from 2015 to 2018, as well as the cluster that included inflammatory diseases and infections, among a few others. The diseases included in these clusters increase with age, and thus, the increase in costs is most likely because of the aging of the population.

Table 7. Trends of the annual costs (relative to all costs) of selected clusters from 2015 to 2018.

Trend and cluster	2015	2016	2017	2018
Increasing trend, %				
Tumors	7	7.1	7.2	7.4
Mixed cluster 10	6.1	6.3	6.4	6.6
Lower respiratory system	6.1	6.2	6.4	6.4
Eye and ear	2.9	3	3.1	3.2
Decreasing trend, %				
Mental and behavioral disorders	9.5	9	9	8.8
Mixed cluster 8	6.9	6.7	6.7	6.3
Injuries	4.4	4.2	4.2	4
Pregnancy	2.4	2.2	2	2

The relative costs of mental and behavioral disorders decreased the most (from 9.5% to 8.8%), whereas injuries (4.4% to 4.0%) and pregnancy-related diseases (2.4% to 2.0%) also showed a clear decrease. There are several explanations for the observed decline in the costs of care related to mental and behavioral disorders, including the current tendency to prefer outpatient services and difficulties in appropriate service provision. The absolute cost values for pregnancy-related issues were €219 million, €213 million, €202 million, and 194 million from 2015 to 2018. Therefore, the decrease is real, which could be explained by the decrease in the birth rate from 1.65 to 1.41 during the same period (1.65, 1.57, 1.49, and 1.41) [64].

Discussion

Principal Findings

We analyzed the data by clustering the diagnoses into 15 clusters. All clusters were consistent with expert knowledge of the domain. Some of these clusters were expected. For example, mental and behavioral disorders were so closely associated with substance abuse problems that they formed one cluster. Some clusters also showed interesting and unexpected connections, such as a cluster that included lower respiratory tract diseases and systemic connective tissue disorders. Although some connections are easily justified by the close relation of the diagnoses, they are not necessarily considered when planning the current service processes and resources. For example, understanding the strong connections between many disorders related to aging could improve the treatment processes of older patients who are multimorbid.

Analysis of the connections between clusters also provided interesting details. For example, the mental health and substance abuse cluster was very closely connected to the cluster comprising fractures and other injuries. A possible explanation is that mental health and substance abuse problems often lead to painful, fracture-causing accidents. The nutritional problems cluster was the most central in the data, with a strong connection to 10 other clusters. This is an interesting finding that addresses the connection between nutritional status and various health disorders.

For each patient, the highest costs were in cluster 2 (€5435), which included infectious diseases that strongly affect the immune system, diseases of the blood and blood-forming organs, and other disorders involving the immune mechanism. These diseases are likely to need frequent contact with specialized care.

Clusters associated with an aging population increased their proportion of all costs from 2015 to 2018. These clusters included diseases related to tumors, lower respiratory system, and eye and ear. The relative costs of mental and behavioral disorders decreased the most (from 9.5% to 8.8%), which might be partly explained by the current tendency to prefer outpatient services.

Limitations

The underlying data reflect how patients use health services and are diagnosed during health care contacts, which may not always accurately reflect the true relationship between diseases. For example, a person who visits health services only for caries treatment may not be as easily diagnosed with alcohol-related disorders (F10) or problems related to metabolic disorders (E66) as a person who visits because of mental health issues or maternity issues.

The clustering methodology itself has a few limitations. Although the chosen clustering algorithm and cost function were shown to have good clustering accuracy with validation data, it forces every diagnosis to belong to a cluster, even if it does not have any connections to other diagnoses. A possible improvement could be the application of outlier detection as a preprocessing step to remove such cases.

Another limitation is that every diagnosis can belong to only one cluster, although it can be connected to diseases in several clusters. For example, dental health diagnoses were clustered with mental retardation and malformations but are clearly very relevant comorbidities for other chronic conditions such as

diabetes. In addition, many infectious disease subgroups are likely to have significant connections with many chronic conditions that decrease the immune response, such as tumors.

The data might also be biased by domestic characteristics within the Finnish population and traditions in recording diagnoses. For example, some conditions such as substance abuse disorders are still highly stigmatized and thus underdiagnosed. The research goal was to find relevant multimorbidity diseases that have a high cost effect on the Finnish health care system. Although some bias might exist, we expect most multimorbidity patterns to appear in other high-income countries, and therefore, the main results might be globally generalizable. This finding was partly confirmed by similar studies in the United States [65] and France [23].

Comparison with other clustering results in earlier studies was challenging mainly because there are many variations in the definition and measures of multimorbidity, as well as the data sources, such as registers, health records, and self-reports, which have been used to obtain information on comorbidities. These differences make comparison difficult but still possible to some degree, as shown in the studies by Prados-Torres et al [13] and Wartelle et al [23].

Comparison With Prior Work

Comparison of Clusters

Wartelle et al [23] obtained 16 clusters (vs 15 in our case). Some of these were similar to ours. For example, cluster 5 contained diagnoses related to mental disorders, substance abuse, and fractures. In our results, substance abuse and mental problems also formed a cluster, which was closely connected to another cluster with different types of fractures. Their data also included one women-specific cluster with pregnancy-related diagnoses. However, most of the clusters were very different from ours.

Their clusters were more unbalanced in size; 5 of the clusters contained only 1 diagnosis, and the largest cluster had 13 diagnoses. In our case, the smallest cluster was size 13, and the largest size was 15. This is partly because of our choice of a clustering cost function that favors more balanced clusters and also because the choice of ED data in [23] was expected to generate larger clusters for trauma diagnoses.

Most of the differences originated from the data. Our data are from everyday health care visits, whereas the data studied by Wartelle et al [23] came from ED visits. They had a smaller number of diagnoses (162 vs 205). These included symptom codes (R00-R99) and factors influencing health status (Z00-Z99), which we removed as we found them to confuse the analysis. These data-related factors produced several clear differences in the results, which we report in the following sections.

The first difference from the study by Wartelle et al [23] is that our data had a female majority (2,062,110/3,835,531, 53.7%). We had only 3 clusters with more male than female patients (nutritional 314,390/469,378, 66.9%; injuries 516,849/1,008,118, 51.2%; immune system and blood-forming organs 110,157/216,898, 50.7%). The ED data had 10 clusters with a male majority (52%-64%). A likely explanation is that these clusters were either directly or indirectly related to trauma commonly treated in EDs, whereas our data represent the services used in primary health care, which has only one cluster (cluster 12) related to injuries.

Patients in the ED data were also much younger than those in our data (mean age 40 years vs 51 years). There were 3 clusters in which the average age of patients exceeded 50 years. One of the clusters (approximately 50%) mostly comprised children aged <5 years. Our data were restricted to adult patients. ED data also lacked a clear pregnancy cluster, and pregnancy-related diagnoses were merged with digestive- and menstruation-related diagnoses.

Busija et al [66] conducted a meta-analysis investigating 51 different articles on multimorbidity profiles. They constructed a similarity matrix of health conditions by counting the number of times each pair of diseases appeared within the same group. The similarity matrix was then projected onto a 2D surface using multidimensional scaling (SPSS/PROXSCAL). This was performed separately for 4 different types of studies grouped by methodology: exploratory factor analysis, cluster analysis of diseases, latent class analysis, and cluster analysis of people.

Overall, their data had fewer diagnoses and clusters. The largest case (factor analysis) included only 70 diagnoses, and they manually distinguished 5 clusters (with a group of mental health problems as one axis) from the 2D projection. They reported clustering of vision, hearing impairment, and fractures in 2 of the 4 cases. In our data, vision and hearing problems were in one cluster, and fractures were in another. These were also weakly connected. A mental health group was visible in all 4 cases and was closely associated with addictions. This is consistent with our results, where mental health and substance abuse problems formed 1 cluster.

Comparison of Costs

We compared the cost of our data with that reported by the Milken Institute in the United States in 2016 [65]. The costliest (both direct and indirect costs) chronic disease in the United States is diabetes type 2, with direct costs of US $185 billion. When indirect costs are included, the four most costly diseases were hypertension (US $1042 billion), diabetes type 2 (US $526 billion), chronic back pain (US $440 billion), and osteoarthritis (US $430 billion).

The costliest diseases (hypertension and type 2 diabetes) are in accordance with our results, where the costliest is cluster 15 (cardiovascular and metabolic), which includes hypertension and diabetes-related diagnoses (I10-I15 and E10-E14), as well as other related cardiovascular diseases common in the Finnish population. The costs of the cluster become high as the size of the patient population increases, as well as the need for frequent contact with health care, although costs per visit are close to average.

Conclusions

To the best of our knowledge, this is the first clustering study with such a rich data set, including all health care visits of Finnish adults aged ≥18 years, covering both primary- and secondary-level care. Good coverage is important, as the

tendency in the development of health service systems is to seek better integration of services, including the integration of primary health care, specialized care, and social services.

Identifying multimorbidity clusters, related characteristics, and especially the burden they cause for service use and costs is helpful in estimating the resources needed in the service system, including the specialties and other knowledge profiles of professionals. Such information could also be applied to estimate future needs when, for example, the projections of population aging and other demographics are known.

To the best of our knowledge, this is the first study to use k-means–based clustering of diseases. Although the standard k-means algorithm can be unstable, we used a recent modification called the M-algorithm, which was shown to be accurate on controlled validation data sets. This directly optimizes a cost function for a network that has RR values as weights. Existing studies rely mainly on agglomerative clustering, using either a heuristic cost function such as average or complete linkage or a slow calculation of the RR. The methodology used was accurate and scalable for large-scale data.

In a future study, we will consider clustering patients and comparing whether the same diagnoses can be grouped together. Another idea is to study geographical differences within Finland. The data are large, and as they are publicly available, they have a high potential for others to find more interesting results by data mining.

Acknowledgments

The project was funded by the Strategic Research Council at the Academy of Finland (grant numbers 312703, 312706, 336325, and 336330). The authors thank Professor Miika Linna for their help with the cost estimations.

Authors' Contributions

SS and PF developed the analytical methods. TL (MD) and KW performed the data acquisition and medical interpretation and provided guidance related to the health service system. PF and SS drafted the manuscript. All the authors contributed to the writing and editing of the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
ICD-10 (International Classification of Diseases, 10th Revision) blocks.
[DOCX File , 29 KB - medinform_v10i5e35422_app1.docx]

References

1. van Oostrom SH, Picavet HS, de Bruin SR, Stirbu I, Korevaar JC, Schellevis FG, et al. Multimorbidity of chronic diseases and health care utilization in general practice. BMC Fam Pract 2014 Apr 07;15:61 [FREE Full text] [doi: 10.1186/1471-2296-15-61] [Medline: 24708798]
2. van den Akker M, Buntinx F, Knottnerus JA. Comorbidity or multimorbidity. Eur J Gen Pract 1996;2(2):65-70. [doi: 10.3109/13814789609162146]
3. van den Akker M, Buntinx F, Metsemakers JF, Roos S, Knottnerus JA. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. J Clin Epidemiol 1998 May;51(5):367-375. [doi: 10.1016/s0895-4356(97)00306-5] [Medline: 9619963]
4. Willadsen TG, Bebe A, Køster-Rasmussen R, Jarbøl DE, Guassora AD, Waldorff FB, et al. The role of diseases, risk factors and symptoms in the definition of multimorbidity - a systematic review. Scand J Prim Health Care 2016 Jun;34(2):112-121 [FREE Full text] [doi: 10.3109/02813432.2016.1153242] [Medline: 26954365]
5. Xu X, Mishra GD, Jones M. Evidence on multimorbidity from definition to intervention: an overview of systematic reviews. Ageing Res Rev 2017 Aug;37:53-68. [doi: 10.1016/j.arr.2017.05.003] [Medline: 28511964]
6. Wang L, Si L, Cocker F, Palmer AJ, Sanderson K. A systematic review of cost-of-illness studies of multimorbidity. Appl Health Econ Health Policy 2018 Feb;16(1):15-29. [doi: 10.1007/s40258-017-0346-6] [Medline: 28856585]
7. Brettschneider C, Leicht H, Bickel H, Dahlhaus A, Fuchs A, Gensichen J, MultiCare Study Group. Relative impact of multimorbid chronic conditions on health-related quality of life--results from the MultiCare Cohort Study. PLoS One 2013;8(6):e66742 [FREE Full text] [doi: 10.1371/journal.pone.0066742] [Medline: 23826124]
8. Stirland LE, González-Saavedra L, Mullin DS, Ritchie CW, Muniz-Terrera G, Russ TC. Measuring multimorbidity beyond counting diseases: systematic review of community and population studies and guide to index choice. BMJ 2020 Feb 18;368:m160 [FREE Full text] [doi: 10.1136/bmj.m160] [Medline: 32071114]
9. Farley JF, Harley CR, Devine JW. A comparison of comorbidity measurements to predict healthcare expenditures. Am J Manag Care 2006 Feb;12(2):110-119 [FREE Full text] [Medline: 16464140]

10. Travers DA, Haas SW, Waller AE, Tintinalli JE. Diagnosis clusters for emergency medicine. Acad Emerg Med 2003 Dec;10(12):1337-1344 [FREE Full text] [doi: 10.1111/j.1553-2712.2003.tb00008.x] [Medline: 14644786]
11. Schneeweiss R, Cherkin DC, Hart LG, Revicki DA, Wollstadt LJ, Stephenson MJ, et al. Diagnosis clusters adapted for ICD-9-CM and ICHPPC-2. J Fam Pract 1986 Jan;22(1):69-72. [Medline: 3079816]
12. Jain AK, Dubes RC. Algorithms for clustering data. Hoboken, NJ, USA: Prentice-Hall; Jan 1988.
13. 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]
14. Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol 2009 Apr;5(4):e1000353 [FREE Full text] [doi: 10.1371/journal.pcbi.1000353] [Medline: 19360091]
15. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak 2019 Jul 23;19(1):142 [FREE Full text] [doi: 10.1186/s12911-019-0852-6] [Medline: 31337390]
16. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform 2019 Nov;99:103291 [FREE Full text] [doi: 10.1016/j.jbi.2019.103291] [Medline: 31560949]
17. Kalgotra P, Sharda R, Croff JM. Examining health disparities by gender: a multimorbidity network analysis of electronic medical record. Int J Med Inform 2017 Dec;108:22-28. [doi: 10.1016/j.ijmedinf.2017.09.014] [Medline: 29132627]
18. Folino F, Pizzuti C, Ventura M. A comorbidity network approach to predict disease risk. In: Proceedings of the 1st International Conference on Information Technology in Bio- and Medical Informatics. 2010 Presented at: ITBAM '10; September 1-2, 2010; Bilbao, Spain p. 102-109. [doi: 10.1007/978-3-642-15020-3_10]
19. Folino F, Pizzuti C. Link prediction approaches for disease networks. In: Proceedings of the 3rd International Conference on Information Technology in Bio- and Medical Informatics. 2012 Presented at: ITBAM '12; September 4-5, 2012; Vienna, Austria p. 99-108. [doi: 10.1007/978-3-642-32395-9_8]
20. Ding R, Jiang F, Xie J, Yu Y. Algorithmic prediction of individual diseases. Int J Prod Res 2017;55(3):750-768. [doi: 10.1080/00207543.2016.1208372]
21. John R, Kerby DS, Hennessy CH. Patterns and impact of comorbidity and multimorbidity among community-resident American Indian elders. Gerontologist 2003 Oct;43(5):649-660. [doi: 10.1093/geront/43.5.649] [Medline: 14570961]
22. Marengoni A, Bonometti F, Nobili A, Tettamanti M, Salerno F, Corrao S, Italian Society of Internal Medicine (SIMI) Investigators. In-hospital death and adverse clinical events in elderly patients according to disease clustering: the REPOSI study. Rejuvenation Res 2010 Aug;13(4):469-477. [doi: 10.1089/rej.2009.1002] [Medline: 20586646]
23. 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]
24. 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]
25. Fränti P, Sieranoja S. How much can k-means be improved by using better initialization and repeats? Pattern Recognition 2019 Sep;93:95-112. [doi: 10.1016/j.patcog.2019.04.014]
26. Sieranoja S, Fränti P. Adapting k-means for graph clustering. Knowl Inf Syst 2021 Dec 04;64(1):115-142. [doi: 10.1007/s10115-021-01623-y]
27. Multimorbidity network analysis. University of Eastern Finland. URL: http://cs.uef.fi/ml/impro/DiagnosisClusters/ [accessed 2022-04-26]
28. Whitty CJ, Watt FM. Map clusters of diseases to tackle multimorbidity. Nature 2020 Mar;579(7800):494-496. [doi: 10.1038/d41586-020-00837-4] [Medline: 32210388]
29. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015 Jan 06;162(1):W1-73 [FREE Full text] [doi: 10.7326/M14-0698] [Medline: 25560730]
30. Newman ME. Modularity and community structure in networks. Proc Natl Acad Sci U S A 2006 Jun 06;103(23):8577-8582 [FREE Full text] [doi: 10.1073/pnas.0601602103] [Medline: 16723398]
31. Newman ME. Analysis of weighted networks. Phys Rev E 2004 Nov 24;70(5):056131. [doi: 10.1103/physreve.70.056131]
32. Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 1970 Feb;49(2):291-307. [doi: 10.1002/j.1538-7305.1970.tb01770.x]
33. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans Pattern Anal Machine Intell 2000 Aug;22(8):888-905. [doi: 10.1109/34.868688]
34. Divo MJ, Casanova C, Marin JM, Pinto-Plata VM, de-Torres JP, Zulueta JJ, BODE Collaborative Group. COPD comorbidities network. Eur Respir J 2015 Sep;46(3):640-650 [FREE Full text] [doi: 10.1183/09031936.00171614] [Medline: 26160874]
35. Hromic H, Prangnawarat N, Hulpus I, Karnstedt M, Hayes C. Graph-based methods for clustering topics of interest in Twitter. In: Proceedings of the 15th International Conference on Engineering the Web in the Big Data Era. 2015 Presented at: ICWE '15; June 23-26, 2015; Rotterdam, The Netherlands p. 701-704. [doi: 10.1007/978-3-319-19890-3_61]
36. Fortunato S. Community detection in graphs. Phys Rep 2010 Feb;486(3-5):75-174. [doi: 10.1016/j.physrep.2009.11.002]

37. Fortunato S, Hric D. Community detection in networks: a user guide. Phys Rep 2016 Nov;659:1-44. [doi: 10.1016/j.physrep.2016.09.002]
38. Blondel VD, Guillaume JL, 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]
39. Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E Stat Nonlin Soft Matter Phys 2009 Jul;80(1 Pt 2):016118. [doi: 10.1103/PhysRevE.80.016118] [Medline: 19658785]
40. Whang JJ, Gleich DF, Dhillon IS. Overlapping community detection using neighborhood-inflated seed expansion. IEEE Trans Knowl Data Eng 2016 May 1;28(5):1272-1284. [doi: 10.1109/tkde.2016.2518687]
41. Lu Z, Wen Y, Cao G. Community detection in weighted networks: algorithms and applications. In: Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications. 2013 Mar 18 Presented at: PerCom '13; March 18-22, 2013; San Diego, CA, USA p. 179-184. [doi: 10.1109/percom.2013.6526730]
42. Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Phys Rev E Stat Nonlin Soft Matter Phys 2009 Nov;80(5 Pt 2):056117. [doi: 10.1103/PhysRevE.80.056117] [Medline: 20365053]
43. Zhang W, Wang X, Zhao D, Tang X. Graph degree linkage: agglomerative clustering on a directed graph. In: Proceedings of the 12th European Conference on Computer Vision. 2012 Presented at: ECCV '12; October 7-13, 2012; Florence, Italy p. 428-441. [doi: 10.1007/978-3-642-33718-5_31]
44. LaSalle D, Karypis G. A parallel hill-climbing refinement algorithm for graph partitioning. In: Proceedings of the 45th International Conference on Parallel Processing. 2016 Presented at: ICPP '16; Aug 16-19, 2016; Philadelphia, PA, USA p. 236-241. [doi: 10.1109/icpp.2016.34]
45. Tabatabaei SS, Coates M, Rabbat M. GANC: greedy agglomerative normalized cut for graph clustering. Pattern Recognit 2012 Feb;45(2):831-843. [doi: 10.1016/j.patcog.2011.06.018]
46. Mustonen E, Hörhammer I, Absetz P, Patja K, Lammintakanen J, Talja M, et al. Eight-year post-trial follow-up of health care and long-term care costs of tele-based health coaching. Health Serv Res 2020 Apr;55(2):211-217 [FREE Full text] [doi: 10.1111/1475-6773.13251] [Medline: 31884682]
47. Linna M, Mikkola T, Peltokorpi A, Tyni T. Rekistereistä tietoa vanhuspalveluiden johtamiseen? Ikääntyneen väestön sosiaali- ja terveyspalveluiden käytön arviointi rekisteriaineistoja hyödyntämällä. Helsinki, Finland: Suomen Kuntaliitto; 2016.
48. Kapiainen S, Väisänen A, Haula T. Terveyden- ja sosiaalihuollon yksikkökustannukset Suomessa vuonna 2011. 2014. URL: https://www.julkari.fi/bitstream/handle/10024/114683/THL_RAPO3_2014_web.pdf?sequence=1&isAllowed=y [accessed 2022-03-23]
49. Srinivasan K, Currim F, Ram S. Predicting high-cost patients at point of admission using network science. IEEE J Biomed Health Inform 2018 Nov;22(6):1970-1977. [doi: 10.1109/JBHI.2017.2783049] [Medline: 29990022]
50. Cornell JE, Pugh JA, Williams Jr JW, Kazis L, Lee AF, Parchman ML, et al. Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database. Appl Multivar Res 2009 Jan 13;12(3):163-182. [doi: 10.22329/amr.v12i3.658]
51. Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of data. 1997 Jun Presented at: SIGMOD '97; May 11-15, 1997; Tucson, AZ, USA p. 265-276. [doi: 10.1145/253260.253327]
52. Moni MA, Liò P. comoR: a software for disease comorbidity risk assessment. J Clin Bioinforma 2014 Mar 23;4:8 [FREE Full text] [doi: 10.1186/2043-9113-4-8] [Medline: 25045465]
53. Dunning AJ, Kensler J, Coudeville L, Bailleux F. Some extensions in continuous models for immunological correlates of protection. BMC Med Res Methodol 2015 Dec 28;15:107 [FREE Full text] [doi: 10.1186/s12874-015-0096-9] [Medline: 26707389]
54. Aguado A, Moratalla-Navarro F, López-Simarro F, Moreno V. MorbiNet: multimorbidity networks in adult general population. Analysis of type 2 diabetes mellitus comorbidity. Sci Rep 2020 Feb 12;10(1):2416 [FREE Full text] [doi: 10.1038/s41598-020-59336-1] [Medline: 32051506]
55. Klimek P, Kautzky-Willer A, Chmiel A, Schiller-Frühwirth I, Thurner S. Quantification of diabetes comorbidity risks across life using nation-wide big claims data. PLoS Comput Biol 2015 Apr;11(4):e1004125 [FREE Full text] [doi: 10.1371/journal.pcbi.1004125] [Medline: 25855969]
56. HuDiNe Search. URL: https://web.archive.org/web/20160426043824/http://hudine.neu.edu/ [accessed 2022-03-23]
57. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the International AAAI Conference on Web and Social Media. 2009 Presented at: ICWSM '09; May 17-20, 2009; San Jose, CA, USA p. 361-362. [doi: 10.1007/978-1-4614-6170-8_299]
58. Rózemberczki B, Davies R, Sarkar R, Sutton C. GEMSEC: graph embedding with self clustering. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2019 Aug Presented at: ASONAM '19; August 27-30, 2019; Vancouver, British Columbia, Canada p. 65-72. [doi: 10.1145/3341161.3342890]
59. Zhao Q, Fränti P. WB-index: a sum-of-squares based index for cluster validity. Data Knowl Eng 2014 Jul;92:77-89. [doi: 10.1016/j.datak.2014.07.008]

60. Al- Zoubi MB, al Rawi M. An efficient approach for computing silhouette coefficients. J Comput Sci 2008 Mar 1;4(3):252-255. [doi: 10.3844/jcssp.2008.252.255]
61. Corrigan P. How stigma interferes with mental health care. Am Psychol 2004 Oct;59(7):614-625. [doi: 10.1037/0003-066X.59.7.614] [Medline: 15491256]
62. Findicator - Mortality from ischaemic heart disease. Statistics Finland. 2020. URL: https://findikaattori.fi/en/83 [accessed 2022-03-23]
63. Karolaakso T, Autio R, Näppilä T, Nurmela K, Pirkola S. Socioeconomic factors in disability retirement due to mental disorders in Finland. Eur J Public Health 2020 Dec 11;30(6):1218-1224 [FREE Full text] [doi: 10.1093/eurpub/ckaa132] [Medline: 32929489]
64. Official Statistics of Finland (OSF): Births. Statistics Finland. 2021. URL: http://www.stat.fi/til/synt/2019/synt_2019_2020-04-24_tie_001_en.html [accessed 2021-06-14]
65. Waters H, Graf M. The costs of chronic disease in the U.S. The Milken Institute. 2018 Aug. URL: https://milkeninstitute.org/sites/default/files/reports-pdf/ChronicDiseases-HighRes-FINAL.pdf [accessed 2022-03-23]
66. Busija L, Lim K, Szoeke C, Sanders KM, McCabe MP. Do replicable profiles of multimorbidity exist? Systematic review and synthesis. Eur J Epidemiol 2019 Nov;34(11):1025-1053. [doi: 10.1007/s10654-019-00568-5] [Medline: 31624969]

Abbreviations

ED: emergency department
ICD: International Classification of Diseases
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification
ICD-10: International Classification of Diseases, 10th Revision
IIW: inverse internal weight
RR: relative risk
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 03.12.21; peer-reviewed by A Wartelle, S Ghaemi, SS Amritphale; comments to author 31.01.22; revised version received 25.02.22; accepted 02.03.22; published 04.05.22.

Please cite as:
Fränti P, Sieranoja S, Wikström K, Laatikainen T
Clustering Diagnoses From 58 Million Patient Visits in Finland Between 2015 and 2018
JMIR Med Inform 2022;10(5):e35422
URL: https://medinform.jmir.org/2022/5/e35422
doi:10.2196/35422
PMID:35507390

Original Paper

Predicting Postoperative Mortality With Deep Neural Networks and Natural Language Processing: Model Development and Validation

Pei-Fu Chen[1,2*], MD; Lichin Chen[3*], PhD; Yow-Kuan Lin[1,4], BSc; Guo-Hung Li[1], MSc; Feipei Lai[1,5,6], PhD; Cheng-Wei Lu[2,7], MD, PhD; Chi-Yu Yang[8,9], MD; Kuan-Chih Chen[1,10], MD, MSc; Tzu-Yu Lin[2,7], MD, PhD

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan
[2]Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[4]Department of Computer Science, Columbia University, New York, NY, United States
[5]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[6]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[7]Department of Mechanical Engineering, Yuan Ze University, Taoyuan, Taiwan
[8]Department of Information Technology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[9]Section of Cardiovascular Medicine, Cardiovascular Center, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[10]Department of Internal Medicine, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[*]these authors contributed equally

Corresponding Author:
Tzu-Yu Lin, MD, PhD
Department of Anesthesiology
Far Eastern Memorial Hospital
No 21, Sec 2, Nanya S Rd
Banciao Dist
New Taipei City, 220216
Taiwan
Phone: 886 2 89667000
Fax: 886 2 89665567
Email: drlin1971@gmail.com

Abstract

Background: Machine learning (ML) achieves better predictions of postoperative mortality than previous prediction tools. Free-text descriptions of the preoperative diagnosis and the planned procedure are available preoperatively. Because reading these descriptions helps anesthesiologists evaluate the risk of the surgery, we hypothesized that deep learning (DL) models with unstructured text could improve postoperative mortality prediction. However, it is challenging to extract meaningful concept embeddings from this unstructured clinical text.

Objective: This study aims to develop a fusion DL model containing structured and unstructured features to predict the in-hospital 30-day postoperative mortality before surgery. ML models for predicting postoperative mortality using preoperative data with or without free clinical text were assessed.

Methods: We retrospectively collected preoperative anesthesia assessments, surgical information, and discharge summaries of patients undergoing general and neuraxial anesthesia from electronic health records (EHRs) from 2016 to 2020. We first compared the deep neural network (DNN) with other models using the same input features to demonstrate effectiveness. Then, we combined the DNN model with bidirectional encoder representations from transformers (BERT) to extract information from clinical texts. The effects of adding text information on the model performance were compared using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Statistical significance was evaluated using $P<.05$.

Results: The final cohort contained 121,313 patients who underwent surgeries. A total of 1562 (1.29%) patients died within 30 days of surgery. Our BERT-DNN model achieved the highest AUROC (0.964, 95% CI 0.961-0.967) and AUPRC (0.336, 95%

CI 0.276-0.402). The AUROC of the BERT-DNN was significantly higher compared to logistic regression (AUROC=0.952, 95% CI 0.949-0.955) and the American Society of Anesthesiologist Physical Status (ASAPS AUROC=0.892, 95% CI 0.887-0.896) but not significantly higher compared to the DNN (AUROC=0.959, 95% CI 0.956-0.962) and the random forest (AUROC=0.961, 95% CI 0.958-0.964). The AUPRC of the BERT-DNN was significantly higher compared to the DNN (AUPRC=0.319, 95% CI 0.260-0.384), the random forest (AUPRC=0.296, 95% CI 0.239-0.360), logistic regression (AUPRC=0.276, 95% CI 0.220-0.339), and the ASAPS (AUPRC=0.149, 95% CI 0.107-0.203).

Conclusions: Our BERT-DNN model has an AUPRC significantly higher compared to previously proposed models using no text and an AUROC significantly higher compared to logistic regression and the ASAPS. This technique helps identify patients with higher risk from the surgical description text in EHRs.

(JMIR Med Inform 2022;10(5):e38241) doi:10.2196/38241

KEYWORDS

bidirectional encoder representations from transformers; deep neural network; natural language processing; postoperative mortality prediction; unstructured text; machine learning; preoperative medicine; anesthesia; prediction model; anesthesiologist; deep learning model; electronic health record; neural network

Introduction

The prevalence of postoperative mortality is 0.5%-2.8 % in patients undergoing elective surgery [1]. The risks are attributable to the patient's condition and can be modulated with adequate evaluation and planning during surgery and anesthesia. Several tools have been developed to predict postoperative mortality, including the American College of Surgeons' (ACS) National Surgical Quality Improvement Program (NSQIP) risk calculator, the American Society of Anesthesiologist Physical Status (ASAPS), the risk quantification index, the risk stratification index, and the preoperative score [2-5]. Although these classification systems consider the patient's general condition and surgery category, preoperative vital signs and laboratory data—which are critical in predicting postoperative mortality—are not typically included [6]. Moreover, a patient's surgical information is commonly written as text in the medical record. Although reading this information helps anesthesiologists evaluate the risk of the surgery, it is difficult to include it in a classification tool. These deficiencies make it challenging to identify the small groups of patients with higher risks. Better tools for predicting postoperative mortality remain under investigation.

Machine learning (ML) is widely applied to medical problems, including for predicting postoperative mortality [6-11]. ML models can automatically predict postoperative mortality using electronic health records (EHRs) before surgery, and they achieve a superior area under the receiver operating characteristic curve (AUROC) than previous methods [6]. To stratify surgery types, previous studies have used the Current Procedural Terminology (CPT) codes or *International Classification of Diseases* (ICD) codes for surgical information [2,6,7,9,12]. These methods are not widely applicable, because the CPT is not implemented worldwide and ICD codes are seldom recorded before surgery. In addition, because this surgical information is written in the medical record by surgeons before surgery, using this text in models may improve the prediction of postoperative mortality.

Compared to structured EHRs, unstructured clinical text requires meaningful concept embeddings to be extracted before model training, making it more challenging [13]. However, including this unstructured text improves the advanced prediction of unfavorable clinical outcomes [14-16]. Bidirectional encoder representations from transformers (BERT) is a contextualized embedding method that preserves the distance of meanings with multihead attention [17]. After pretrained on the relevant corpora and proper architecture modification, BERT extracts meaningful embeddings from clinical text [18,19].

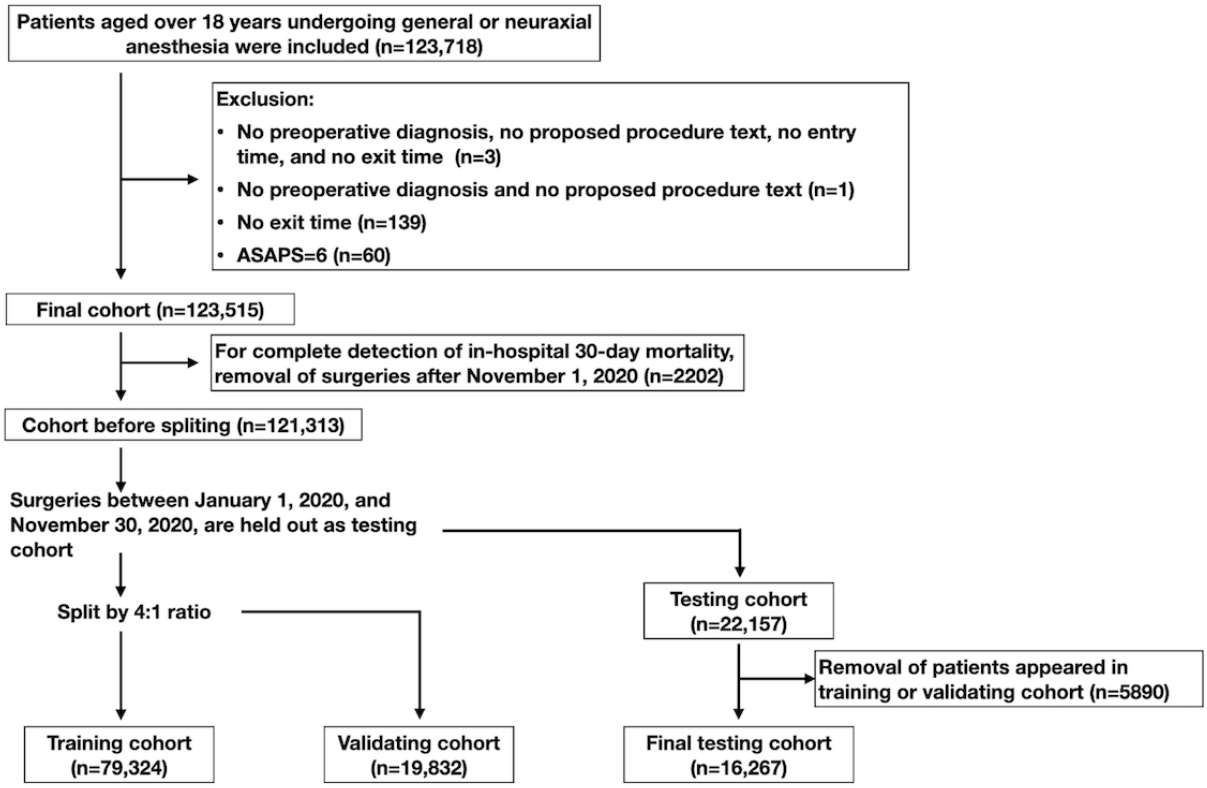
This study aims to develop a model to predict 30-day postoperative mortality before surgery that performs better than state-of-the-art models. Our contribution is including free (ie, unstructured) text in postoperative mortality prediction by proposing a deep neural network (DNN) model with BERT. We investigate the effectiveness of unstructured clinical texts (eg, preoperative diagnosis and proposed procedures) in predicting postoperative mortality.

Methods

Data Extraction

This study aims to predict in-hospital 30-day postoperative mortality using preoperative anesthesia assessments. Data were collected from the electronic health system of the Far Eastern Memorial Hospital, a large academic medical center in Taiwan. Preoperative anesthesia assessment records and discharge summaries were included. Overall, 5 years' worth of retrospective data were collected from January 1, 2016, to December 30, 2020. The last version of the anesthesia assessment was included for each surgery. Patients over 18 years of age who underwent at least 1 surgical procedure under general or neuraxial anesthesia were included. Cases with an ASAPS of 6 were excluded. Records lacking entry time, exit time, preoperative diagnosis, or proposed procedure text were excluded. The in-hospital 30-day postoperative mortality was defined by a discharged route of "expired" and "critical against-advice discharge" (when the patient wants to die at home) without future admission. Discharges within 30 days after surgery were identified and labeled as "true"; those occurring outside this window were marked as "false." The end date of the testing set was November 30, 2020, 30 days before the end of the collected data, to ensure complete 30-day mortality detection (Figure 1).

Figure 1. Flow diagram. ASAPS: American Society of Anesthesiologist Physical Status.



Ethical Approval

The Institutional Review Board of the Far Eastern Memorial Hospital approved this retrospective study and waived the requirement of informed consent (#109129-F and #110028-F).

Data Description

We collected 123,718 surgery results for patients aged over 18 years. After applying the exclusion criteria, a cohort of 123,515 (99.8%) patients who underwent surgeries remained. A final cohort of 121,313 (98.2%) patients was used after removing those who underwent surgeries after November 30, 2020 (Figure 1). The training, validation, and testing cohorts finally contained 79,324 (68.7%), 19,832 (17.2%), and 16,267 (14.1%) of 115,423 patients. Patient characteristics of the training, validation, and testing cohorts are listed in Table 1. In the overall cohort, most patients had an ASAPS of 2 or 3. Overall, 107,176 (88.5%) of patients were under general anesthesia. The most prevalent comorbidities were hypertension (n=43,391, 35.8%), followed by diabetes (n=24,314, 20.0%). A total of 1562 (1.3%), 997 (1.3%), 249 (1.3%), and 215 (1.3%) patients died within 30 days of surgery in the overall, training, validation, and testing cohorts, respectively. Multimedia Appendix 1 present a summary of the laboratory data and preoperative vital signs.

Table 1. Characteristics of the cohort. Categorical variables are represented as frequency (%). Continuous variables are represented as the median (25th, 75th percentile). The testing cohort was split by time between the training and validation cohorts, and those cases arising from the training and validation cohorts were removed to prevent data leakage (n=5890, 4.9%).

Feature	Training cohort (N=79,324)	Validation cohort (N=19,832)	Testing cohort (N=16,267)	Overall cohort (N=121,313)
Age (years), median (25th, 75th percentile)	54 (40, 66)	54 (40, 66)	53 (39, 65)	55 (41, 66)
Male sex, n (%)	40,444 (51.0)	9922 (50.0)	8101 (49.8)	61,485 (50.7)
Height (cm), median (25th, 75th percentile)	162 (157, 168)	162 (156, 168)	162 (157, 169)	162 (157, 168)
Weight (kg), median (25th, 75th percentile)	64 (56, 74)	64 (56, 74)	65 (56, 75)	64 (56, 74)
BMI, median (25th, 75th percentile)	24 (22, 27)	24 (22, 27)	24 (22, 27)	24 (22, 27)
ASAPS[a], n (%)				
1	2925 (3.7)	739 (3.7)	660 (4.1)	4404 (3.6)
2	54,056 (68.15)	13,549 (68.3)	11,508 (70.7)	82,588 (68.1)
3	20,842 (26.3)	5155 (26.0)	,654 (22.5)	31,878 (26.3)
4	1345 (1.70)	355 (1.8)	397 (2.4)	2204 (1.8)
5	156 (0.2)	34 (0.2)	48 (0.3)	239 (0.2)
ASA[b] emergency, n (%)	6379 (8.0)	1615 (8.1)	1678 (10.3)	9942 (8.2)
Anesthesia type, n (%)				
General	69,898 (88.3)	17,497 (88.4)	14,486 (89.2)	107,176 (88.5)
Neuraxial	9297 (11.7)	2303 (11.6)	1748 (10.8)	13,929 (11.5)
Emergency level of surgery, n (%)				
Elective	62,226 (78.5)	15,455 (77.9)	12,000 (73.8)	94,816 (78.2)
Urgent	13,800 (17.4)	3,567 (18.0)	3,356 (20.6)	21,342 (17.6)
Emergency	2849 (3.6)	708 (3.6)	801 (4.9)	4484 (3.7)
Immediate	449 (0.57)	102 (0.51)	110 (0.7)	671 (0.6)
Preoperative location, n (%)				
Ward	47,187 (59.5)	11,788 (59.4)	9824 (60.4)	72,045 (59.4)
Outpatient	18,386 (23.2)	4463 (22.5)	2995 (18.4)	27,830 (22.9)
Emergency department	10,083 (12.7)	2592 (13.1)	2283 (14.0)	15,247 (12.6)
Intensive care unit	3668 (4.6)	989 (5.0)	1165 (7.2)	6191 (5.1)
Surgery department, n (%)				
Urology	14,760 (18.6)	3630 (18.3)	2665 (16.4)	22,471 (18.5)
General	11,416 (14.4)	2926 (14.8)	2457 (15.1)	17,608 (14.5)
Orthopedics	10,976 (13.8)	2748 (13.9)	2338 (14.4)	16,772 (13.8)
Gynecology[c]	10,206 (12.9)	2,578 (13.0)	2,302 (14.2)	15,679 (12.9)
Cardiovascular	8692 (11.0)	2086 (10.5)	1491 (9.2)	13,049 (10.8)
Otolaryngology	6193 (7.8)	1505 (7.6)	1223 (7.5)	9427 (7.8)
Plastic surgery	5116 (6.5)	1294 (6.5)	1077 (6.6)	7821 (6.4)
Neurosurgery	3233 (4.1)	833 (4.2)	727 (4.5)	4955 (4.1)
Traumatology	2808 (3.5)	740 (3.7)	722 (4.4)	4357 (3.6)
Thoracic surgery	2006 (2.5)	514 (2.6)	430 (2.6)	3104 (2.6)
Colorectal surgery	1679 (2.1)	423 (2.1)	331 (2.0)	2574 (2.1)
Others	2239 (2.8)	555 (2.8)	504 (3.1)	3496 (2.9)
Comorbidity, n (%)				

Feature	Training cohort (N=79,324)	Validation cohort (N=19,832)	Testing cohort (N=16,267)	Overall cohort (N=121,313)
Diabetes mellitus	15,906 (20.1)	3863 (19.5)	2812 (17.3)	24,314 (20.0)
Hyperlipidemia	8704 (11.0)	2119 (10.7)	1740 (10.7)	13,678 (11.3)
Hypertension	28,462 (35.9)	7055 (35.6)	4999 (30.7)	43,391 (35.8)
Prior cerebrovascular accident	4355 (5.5)	1028 (5.2)	717 (4.4)	6564 (5.4)
Cardiac disease	13,215 (16.7)	3254 (16.4)	2227 (13.7)	20,156 (16.6)
Chronic obstructive pulmonary disease	1549 (2.0)	380 (1.9)	286 (1.8)	2428 (2.0)
Asthma	3024 (3.8)	762 (3.8)	592 (3.6)	4626 (3.8)
Hepatic disease	9118 (11.5)	2299 (11.6)	1664 (10.2)	13,887 (11.4)
Renal disease	12,471 (15.7)	3095 (15.6)	1466 (9.0)	18,874 (15.6)
Bleeding disorder	11,243 (14.2)	2684 (13.5)	2122 (13.0)	17,543 (14.5)
Prior major operations	54,356 (68.5)	13,592 (68.5)	10,040 (61.7)	83,490 (68.8)
Smoking	20,235 (25.5)	5098 (25.7)	3719 (22.9)	30,433 (25.1)
Drug allergy	11,662 (14.7)	2959 (14.9)	2190 (13.5)	18,092 (14.9)
Consciousness	69,858 (88.1)	17,461 (88.0)	15,107 (92.9)	107,906 (88.9)
30-day mortality, n (%)	997 (1.3)	249 (1.3)	215 (1.3)	1562 (1.3)

[a]ASAPS: American Society of Anesthesiologist Physical Status.

[b]ASA: American Society of Anesthesiologists.

[c]The gynecology department consists of gynecology and obstetrics.

Data Preparation

The input features included patient characteristics (age, height, weight, BMI, sex, ASAPS, ASA emergency status, department, preoperative location, and anesthesia type), surgery characteristics (emergency level, preoperative diagnosis, and proposed procedure), comorbidities (diabetes mellitus, hyperlipidemia, hypertension, cerebrovascular accident, cardiac disease, chronic obstructive pulmonary disease, asthma, hepatic disease, renal disease, bleeding disorder, major operations, smoking, and drug allergy), preoperative laboratory data (hemoglobin, platelet, international normalized ratio, prothrombin time, activated partial thromboplastin time, creatinine, aspartate transaminase, alanine transaminase, blood sugar, serum sodium, and serum potassium), and preoperative vital signs (body temperature, oxygen saturation, heart rate, respiratory rate, systolic and diastolic blood pressure, and consciousness status); see Table 2.

Continuous features (eg, age, height, weight, latest laboratory data before surgery, and preoperative vital signs) were standardized by subtracting the mean and scaling to variance. Outliers were regarded as input errors and treated as missing data. Multimedia Appendix 2 lists the definitions of the outliers. Missing values were imputed with the median value of the data set for continuous features.

Categorical features with only 2 classes (eg, sex, comorbidities, ASA emergency status, and consciousness status) were converted into binary encoding. All other categorical features (eg, ASAPS [5 classes], department [22 classes], emergency level [4 classes], preoperative location [4 classes], and anesthesia type [4 classes]) were transformed into one-hot encodings. Missing data were imputed with the majority category of the training data set. The preoperative diagnoses and proposed procedures were expressed as free text. Characters other than alphabetical and numerical ones were removed (eg, Chinese characters [typically notes for colleagues only] and punctuation). English stop words providing no helpful information to the model (eg, “a,” “in,” and “the”) were removed using the Natural Language Toolkit [20].

We used the previous 4 years’ surgery results to predict the last year results. Patients who underwent surgeries between January 1, 2016, and December 31, 2019, were selected and split into training and validation sets in a 4:1 ratio; those who underwent surgeries between January 1, 2020, and November 30, 2020, were selected as the testing set (Figure 1). Patients in the training or validation set were removed from the testing set to prevent information leakage [6].

Table 2. Feature groups included in the models.

Feature type	Feature classes[a]
Patient characteristics	
Continuous	Age, height, weight, BMI
Categorical	Sex (2), ASAPS[b] (5), ASA[c] emergency (2), department (22), preoperative location (4), anesthesia type (4)
Surgery characteristics	
Categorical	Emergency level (4)
Free text	Preoperative diagnosis, proposed procedure
Comorbid conditions	
Categorical	Diabetes mellitus (2), hyperlipidemia (2), hypertension (2), cerebrovascular accident (2), cardiac disease (2), chronic obstructive pulmonary disease (2), asthma (2), hepatic disease (2), renal disease (2), bleeding disorder (2), major operations (2), smoking (2), drug allergy (2)
Preoperative laboratory values	
Continuous	Hemoglobin, platelet, international normalized ratio, prothrombin time, activated partial thromboplastin time, creatinine, aspartate transaminase, alanine transaminase, blood sugar, serum sodium, serum potassium
Preoperative vital signs	
Continuous	Body temperature, oxygen saturation, heart rate, respiratory rate, systolic and diastolic blood pressure
Categorical	Consciousness status (2)

[a]The number of classes is shown in parentheses.

[b]ASAPS: American Society of Anesthesiologist Physical Status.

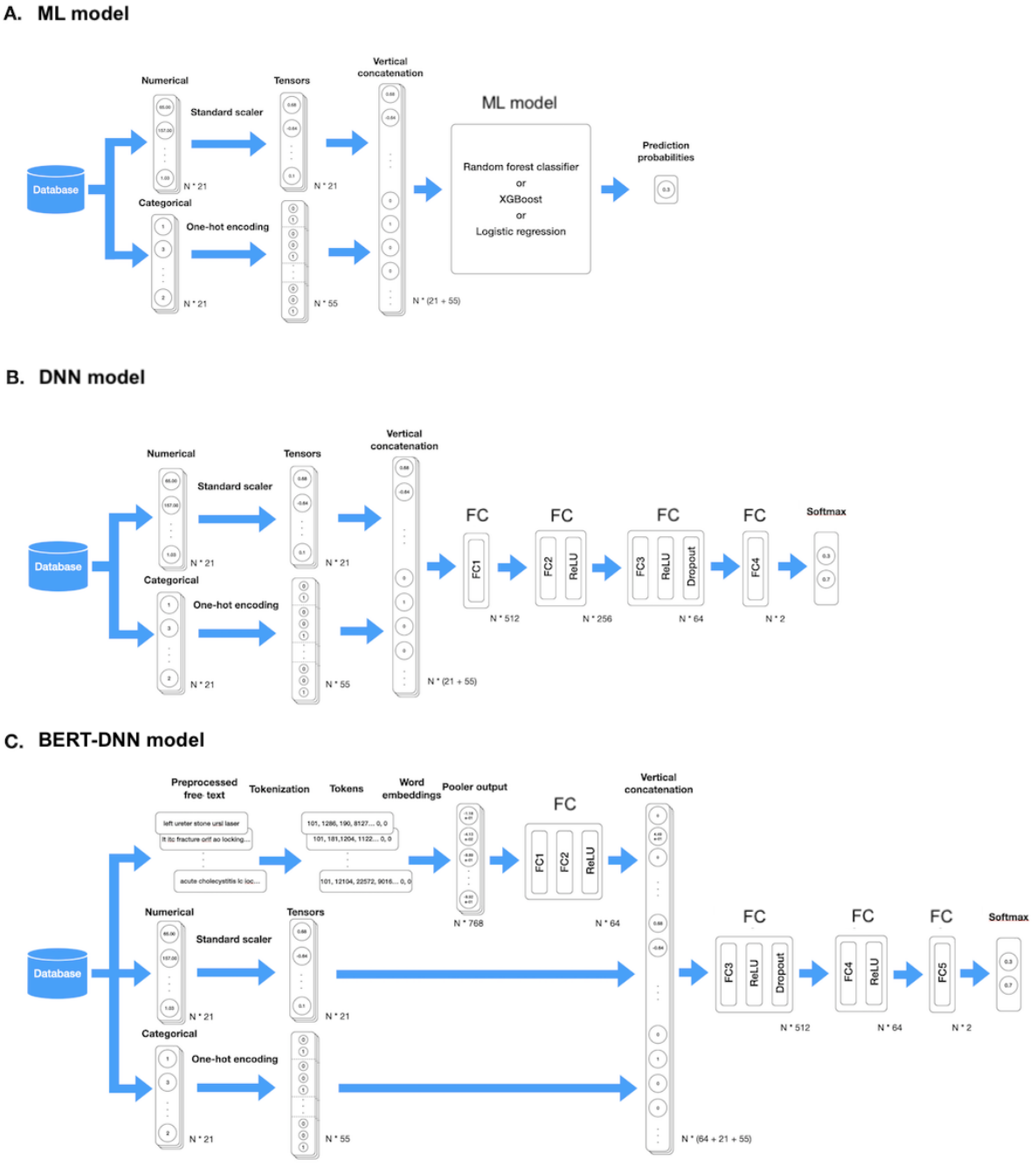
[c]ASA: American Society of Anesthesiologists.

Study Design

Our results were compared with state-of-the-art models, using patient preoperative vital signs and laboratory data to predict in-hospital 30-day mortality [6]. Meanwhile, to demonstrate the effect of adding preoperative diagnoses and proposed procedures to the prediction model, we added text features and compared the performances of the highest-performing models.

First, we compared the state-of-the-art models using patient and surgery characteristics (without text), comorbidities, preoperative vital signs, and laboratory data to predict the in-hospital 30-day mortality. Figure 2B shows our proposed DNN model with 4 fully connected (FC) layers and a Softmax layer output function. We compared our DNN model with other ML models, including a random forest classifier (with 2000 estimators and Gini impurity as the splitting criterion) [21], extreme gradient boosting (XGBoost, with a learning rate of 0.3 and a maximum depth of 6) [22], and logistic regression (with an L2 penalty); see Figure 2A. To balance the data while training the ML models, oversampling by 78 times was performed on the training set via the synthetic minority oversampling technique; this produced synthetic samples along a straight line between randomly selected samples in the feature space [23]. While training our DNN model, we adjusted the weight to compensate for the imbalanced classes. We added the text of preoperative diagnoses and proposed procedures to the DNN model architecture (denoted as BERT-DNN; see Figure 2C) and compared its performance with those of other models.

Figure 2. Architectures of models. BERT: bidirectional encoder representations from transformers: DNN: deep neural network; FC: fully connected; ML: machine learning; ReLU: rectified linear unit; XGBoost: extreme gradient boosting.



Language Model and BERT-DNN Model Design

The language model extracted features from the preprocessed text. Figure 2C shows the architecture of the language model. The preprocessed texts were tokenized using the BERT tokenizer, which transformed each word fragment into a unique token designed for use in BERT's pretraining process [17]. Then, these tokens were embedded by Bio+Clinical BERT, a variant of BERT pretrained on text from PubMed and Medical Information Mart for Intensive Care III [24]. The text information was transformed into a 768-dimension vector (the "word embeddings") at the pooler output layer [17,24]. These word embeddings were input into 2 FC layers before concatenation with other structured features. The concatenated vectors were input into 3 FC layers and a Softmax layer output function. Figure 2C shows the architecture of the BERT-DNN model.

Cross-entropy was used as the loss function. Class weight imbalances were compensated for by setting the weights as the inverses of the different classes' frequencies (~1:78). Further, the training data were split into training and validating sets in a 4:1 ratio to train the deep learning (DL) model. We used AdamW from the PyTorch package as the optimizer, setting a learning rate of 0.00002 for both DL models. We trained our BERT-DNN and DNN models with batch sizes of 64 and 512, respectively, until the 100th epoch. The DL model with the

smallest validation loss was selected for performance comparison.

Model Evaluation

The models were evaluated using the AUROC, the area under the precision-recall curve (AUPRC), sensitivity (also referred to as recall), specificity, precision (also called the positive predictive value), and the F1 score. The F1 score was a harmonic mean of recall and precision and was calculated as 2/[(1/recall) + (1/precision)]. Because postoperative mortalities accounted for 1.3% (1562/121,313) of our data set, classes were extremely imbalanced between the positive and negative groups. Here, the AUPRC (which calculated the average precision) was better than the AUROC for evaluating the discrimination of models [25,26]. For comparison of AUROCs, we applied a nonparametric approach proposed by DeLong et al [27] to calculate the SE of the area and the *P* value. $P<.05$ was regarded statistically significant. We calculated exact binomial 95% CIs for the AUROC. For comparison of AUPRCs, we performed bootstrapping 1000 times in the testing set to calculate the difference in areas and the 95% CI [28]. If the 95% CI for the difference in areas does not include 0, it can be concluded that these 2 areas are significantly different ($P<.05$). We performed bootstrapping 1000 times in the testing set to calculate the 95% CI for other metrics [6]. The predicted probabilities were calibrated using the histogram bins technique, using the same observed mortality in each bin of the validation set [8]. After calibration, the mean observed incidences of mortality were plotted against the mean predicted probabilities within groups in the testing set.

Visualization of Word Embeddings

To show the correlation between increased prediction probabilities and text inputs, the *t* distributed stochastic neighbor embedding (SNE) was implemented by reducing the 768 dimensions of the language model's pool output to 2 into a plane [29,30]. Thus, we showed the clustering of word embeddings using assorted colors for different predicted probabilities and different icons for observed mortalities. We randomly resampled 10,000 and 5000 patients who underwent surgeries in the training and testing sets, respectively, to construct this visualization. The language-model-predicted probabilities and observed mortalities for randomly selected text inputs were calculated and listed.

The study was implemented using Python 3.9, Scikit-learn 0.24 [31], imbalanced-learn 0.8.0 [23], PyTorch 1.8 [32], and transformers 4.9 (Hugging Face) [24]. Our models were trained and validated on the NVIDIA Tesla P100-PCIE-16GB graphics processing unit (GPU). The statistical significances of AUROCs and AUPRCs were calculated using MedCalc software (Ostend, Belgium).

Results

Comparison of Machine Learning Models

The BERT-DNN had the highest AUROC of 0.964 (95% CI 0.961-0.967) and the highest AUPRC of 0.336 (95% CI 0.276-0.402); see Table 3 and Figure 3. The random forest achieved the second-highest AUROC of 0.961 (95% CI 0.958-0.964), and the DNN achieved the second-highest AUPRC of 0.319 (95% CI 0.260-0.384). The BERT-DNN model had the highest F1 score of 0.347 (95% CI 0.305-0.388).

The BERT-DNN had a significantly higher AUROC compared to XGBoost, logistic regression, and ASAPS but not a significantly higher AUROC compared to the DNN and the random forest (Table 4). The BERT-DNN also had a significantly higher AUPRC compared to the DNN, random forest, XGBoost, logistic regression, and ASAPS (Table 5).

In the BERT-DNN model, when the predicted probability of mortality increased from 0.2% to 39.4%, the observed incidence increased from 0.2% to 42.7% (Figure 4).

Table 3. Prediction performances of ML[a] models and ASAPS[b] on the testing cohort with 95% CIs.

Model	AUROC[c] (95% CI)	AUPRC[d] (95% CI)	Accuracy[e] (95% CI)	Sensitivity[e] (95% CI)	Specificity[e] (95% CI)	Precision[a] (95% CI)	F1 score[e] (95% CI)
BERT[f]-DNN[g]	0.964 (0.961-0.967)	0.336 (0.276-0.402)	0.955 (0.952-0.958)	0.749 (0.689-0.805)	0.958 (0.955-0.961)	0.193 (0.166-0.219)	0.307 (0.269-0.342)
DNN	0.959 (0.956-0.962)	0.319 (0.260-0.384)	0.913 (0.909-0.917)	0.885 (0.841-0.926)	0.913 (0.909-0.918)	0.120 (0.104-0.136)	0.212 (0.187-0.236)
Random forest	0.961 (0.958-0.964)	0.296 (0.239-0.360)	0.986 (0.984-0.988)	0.167 (0.122-0.222)	0.997 (0.996-0.998)	0.445 (0.341-0.557)	0.242 (0.182-0.314)
XGBoost[h]	0.950 (0.946-0.953)	0.281 (0.225-0.345)	0.986 (0.984-0.987)	0.195 (0.144-0.249)	0.996 (0.995-0.997)	0.409 (0.312-0.500)	0.263 (0.201-0.326)
Logistic regression	0.952 (0.949-0.955)	0.276 (0.220-0.339)	0.904 (0.900-0.909)	0.833 (0.780-0.882)	0.905 (0.901-0.910)	0.105 (0.091-0.119)	0.187 (0.164-0.210)
ASAPS	0.892 (0.887-0.896)	0.149 (0.107-0.203)	0.970 (0.968-0.973)	0.409 (0.342-0.478)	0.978 (0.975-0.980)	0.197 (0.160-0.235)	0.266 (0.220-0.310)

[a]ML: machine learning.

[b]ASAPS: American Society of Anesthesiologist Physical Status.

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

[d]AUPRC: area under the precision-recall curve.

[e]These metrics were calculated without adjusting the threshold (using 0.5 as the cut-off).

[f]BERT: bidirectional encoder representations from transformers.

[g]DNN: deep neural network.

[h]XGBoost: extreme gradient boosting.

Figure 3. Comparison of discrimination of different models. (A) AUROC. (B) AUPRC. ASAPS: American Society of Anesthesiologist Physical Status; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; BERT: bidirectional encoder representations from transformers; DNN: deep neural network; XGBoost: extreme gradient boosting.

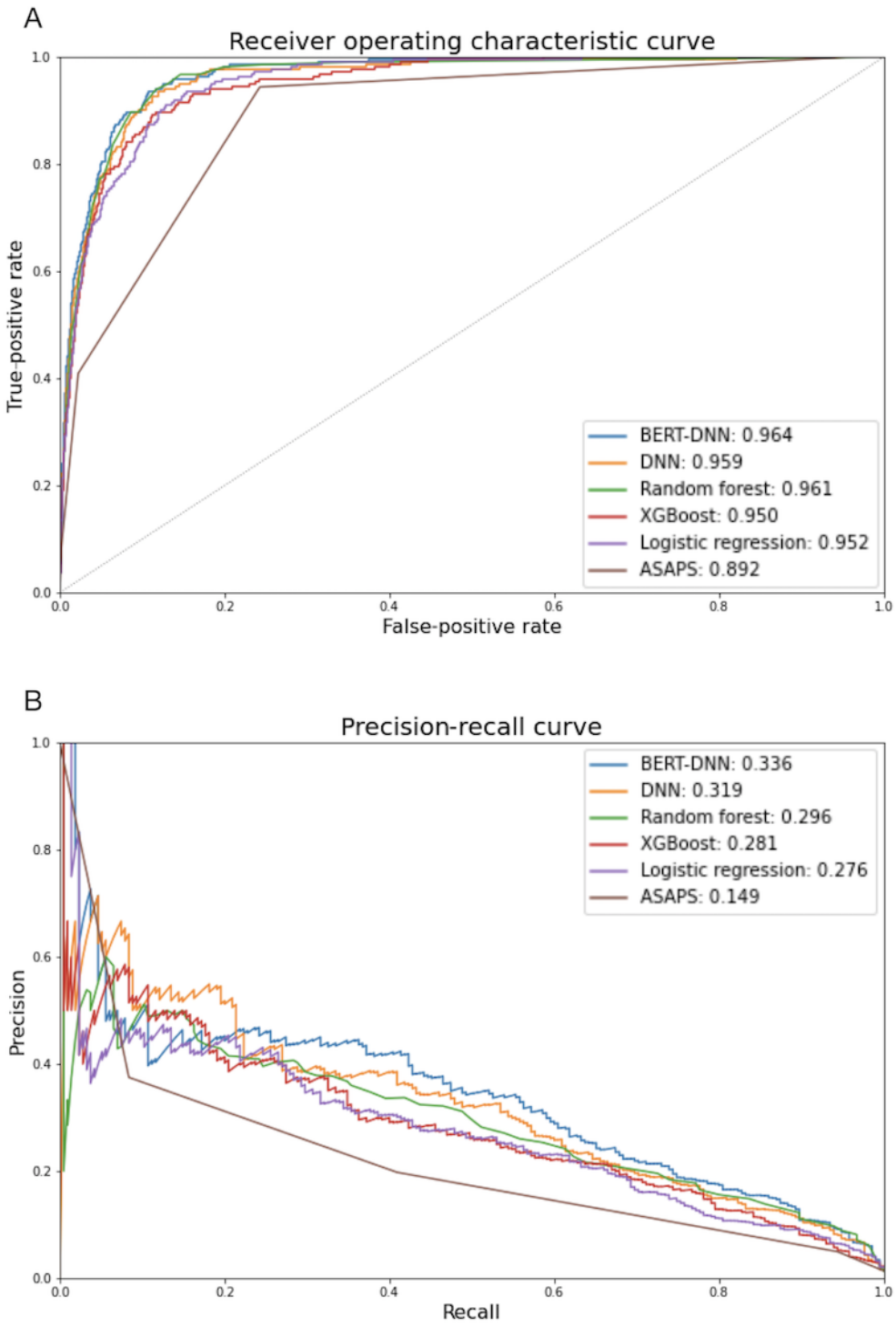


Table 4. Statistical significances of AUROCs[a] of different models. Values are *P* values. We applied a nonparametric approach proposed by DeLong et al [27] to calculate the SE of the area and the *P* value.

	BERT[b]-DNN[c]	DNN	Random forest	XGBoost[d]	Logistic regression
ASAPS[e]	<0.0001[f]	<0.0001[f]	<0.0001[f]	<0.0001[f]	<0.0001[f]
Logistic regression	0.0005[f]	0.0711	0.0351[f]	0.6451	N/A[g]
XGBoost	0.0025[f]	0.0939	0.0262[f]	N/A	N/A
Random forest	0.3816	0.5972	N/A	N/A	N/A
DNN	0.0944	N/A	N/A	N/A	N/A

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

[b]BERT: bidirectional encoder representations from transformers.

[c]DNN: deep neural network.

[d]XGBoost: extreme gradient boosting.

[e]ASAPS: American Society of Anesthesiologist Physical Status.

[f]The difference in areas achieved statistical significance (*P*<.05).

[g]N/A: not applicable.

Table 5. Statistical significances of AUPRCs[a] of different models. Values are differences in areas with 95% CIs calculated by bootstrapping 1000 times [28]. If the 95% CI for the difference in areas does not include 0, it can be concluded that these 2 areas are significantly different (*P*<.05).

	BERT[b]-DNN[c], difference in areas (95% CI)	DNN, difference in areas (95% CI)	Random forest, difference in areas (95% CI)	XGBoost[d], difference in areas (95% CI)	Logistic regression, difference in areas (95% CI)
ASAPS[e]	0.188 (0.159-0.221)[f]	0.170 (0.137-0.201)[f]	0.147 (0.122-0.177)[f]	0.133 (0.107-0.162)[f]	0.127 (0.101-0.154)[f]
Logistic regression	0.061 (0.051-0.073)[f]	0.043 (0.021-0.056)[f]	0.020 (0.006-0.031)[f]	0.006 (–0.006 to 0.014)	N/A[g]
XGBoost	0.055 (0.044-0.068)[f]	0.038 (0.024-0.046)[f]	0.015 (0.005-0.022)[f]	N/A	N/A
Random forest	0.040 (0.030-0.054)[f]	0.023 (0.010-0.032)[f]	N/A	N/A	N/A
DNN	0.018 (0.008-0.037)[f]	N/A	N/A	N/A	N/A

[a]AUPRC: area under the precision-recall curve.

[b]BERT: bidirectional encoder representations from transformers.

[c]DNN: deep neural network.

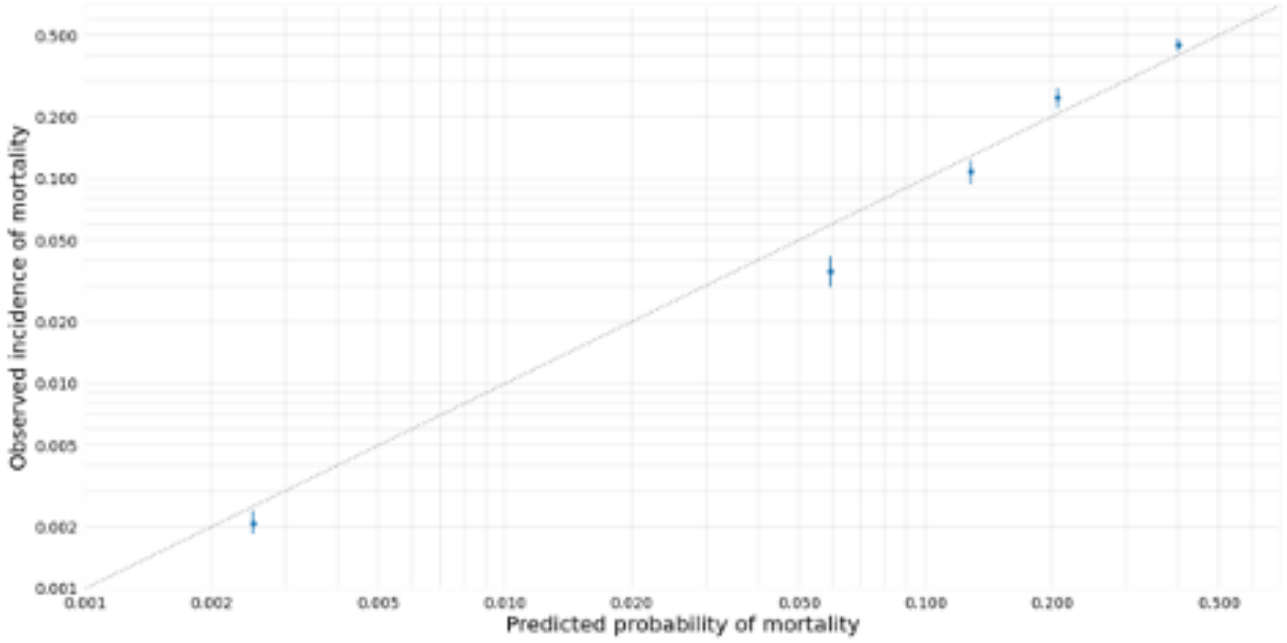
[d]XGBoost: extreme gradient boosting.

[e]ASAPS: American Society of Anesthesiologist Physical Status.

[f]The difference in areas achieved statistical significance (*P*<.05).

[g]N/A: not applicable.

Figure 4. Calibration plot. The observed incidence of mortality was plotted against the calibrated predicted probability of mortality among patients in the test cohort (n=16,267, 14.1%). Predicted probabilities were calibrated by applying the histogram binning technique in the validation cohort using 5 bins. Mean predicted probabilities of in-hospital 30-day mortality were calculated within each group.



Visualization of Word Embeddings

Because the observed mortalities were distributed concordantly with increased prediction probabilities, the annotated scatter plots showed that the text contributed to low- and high-probability predictions (Figure 5, Multimedia Appendix 3). Table 6 lists the probabilities predicted by the language model and the mortalities observed for a randomly selected text input.

Figure 5. Word embeddings visualized by t distributed stochastic neighbor embedding. (A) Word embeddings of the training set. (B) Word embeddings of the testing set. “Probs” indicates probabilities predicated by the BERT-DNN model. The intensity of color increased with the probability. “Labels” indicates mortalities by “x” and survivors by “•”. ards: acute respiratory distress syndrome; atfl: anterior talofibular ligament; avg: arteriovenous graft; avp: aortic valvuloplasty; BERT: bidirectional encoder representations from transformers; bct: breast-conserving therapy; bil: bilateral; bph: benign prostate hypertrophy; bx: biopsy; chr: chronic hypertrophic rhinitis; cps: chronic paranasal sinusitis; dbj: double J stent; DNN: deep neural network; ecmo: extracorporeal membrane oxygenation; emh: endometrial hemorrhage; esrd: end-stage renal disease; fess: functional endoscopic sinus surgery; itc: intertrochanter; ivg: intravenous general anesthesia; lih: left inguinal hernia; mvr: mitral valve replacement; nsd: nasal septum deviation; p: post; pcnl: percutaneous nephrolithotomy; perm cath: permanent catheter; psa: prostate-specific antigen; r: rule out; r’t: right; rirs: retrograde intrarenal surgery; rv: right ventricle; slnd: sentinel lymph node dissection; SNE: stochastic neighbor embedding; t colon: transverse colon; tee: transesophageal echocardiography; tep: total extraperitoneal approach; trus: transrectal ultrasound; turp: transurethral resection of the prostate; urs: ureteroscopy; vats: video-assisted thoracic surgery; vhd: valvular heart disease. Higher-resolution version of this figure available in Multimedia Appendix 3.

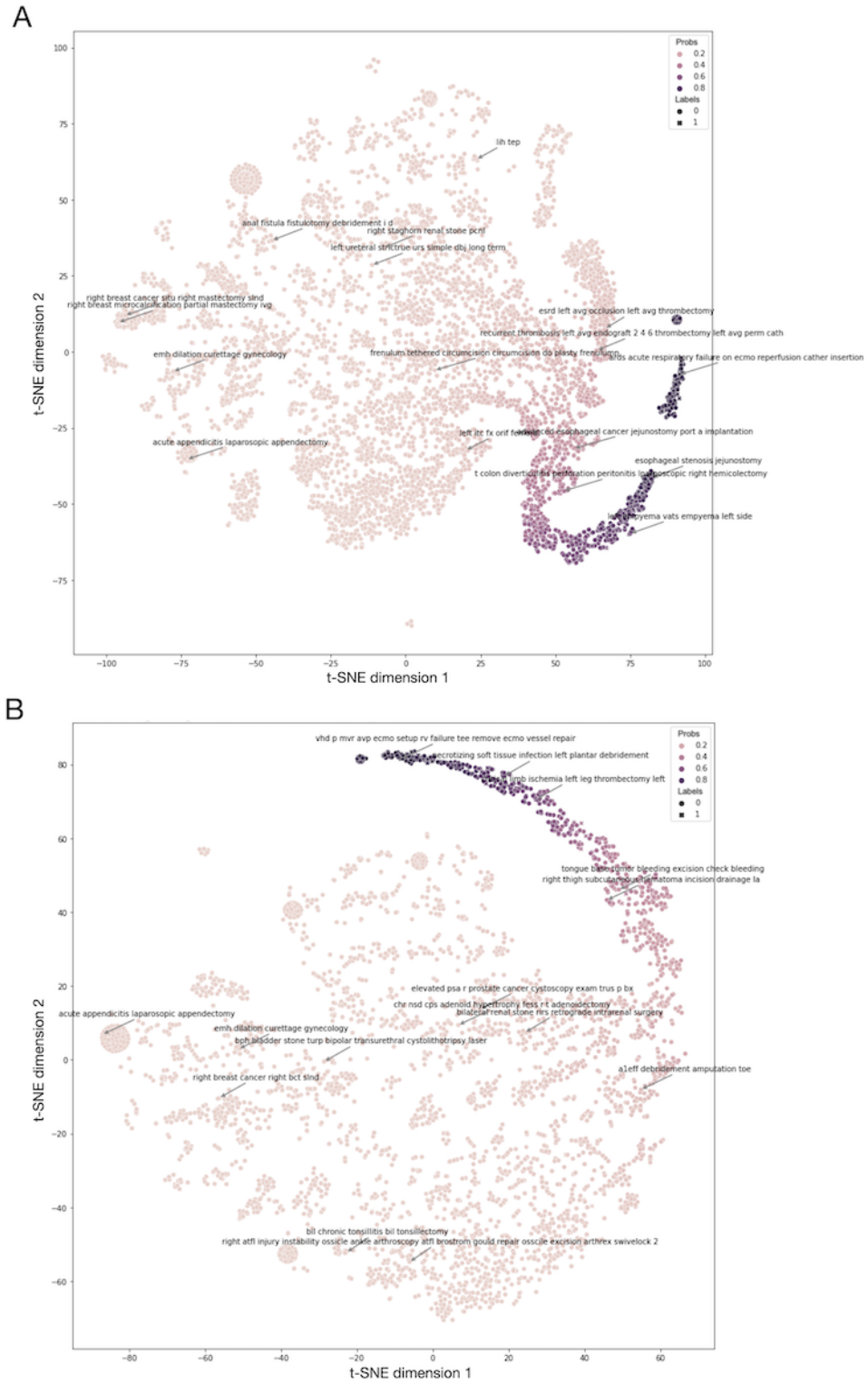


Table 6. Texts and their predicted probabilities by language model. Values are probabilities or mortalities.

Predicted probability	Observed mortality (1=mortality; 0=no mortality)	Original free text combining preoperative diagnosis and proposed procedures
0.951	1	IHCA[a] p[b] CPR[c] ECMO[d] ACS[e] AR[f] full sternotomy CABG[g] AVR[h]
0.948	0	AMI[i] cardiogenic shock p ECMO remove ECMO TEE[j]
0.940	1	hollow organ perforation r[k] PPU[l] related LPPU[m] possible EXP LAP[n]
0.936	1	intra-abdominal bleeding EXP LAP
0.932	1	ischemic bowel laparoscopic diagnosis possible EXP LAP
0.927	0	acute pulmonary embolism IHCA p ECMO angiography TEE
0.925	1	duodenal ulcer perforation p duodenorrhaphy leakage bleeding EXP LAP
0.912	0	respiratory failure tracheostomy
0.880	0	hallow organ perforation r PPU LPPU
0.815	0	acute kidney failure perm cath[o] insertion
0.760	0	post UPPP[p] wound bleeding check bleeding
0.680	0	ESRD[q] HD[r] via right perm cath[s] qw2 4 6 perm cath dysfunction perm cath insertion change perm cath right neck
0.527	0	ESRD left AVG[t] occlusion left AVG thrombectomy
0.415	0	left lower leg soft tissue infection suspect necrotizing fasciitis debridement
0.353	0	ESRD right AVF[u] dysfunction upper arm angiography PTA[v]
0.250	0	RLL[w] lung tumor r lung cancer vats RLL lobectomy wedge first send frozen exam
0.186	0	left lower extremity NF[x] open BK[y]
0.114	0	left anterior mediastinal tumor multiple lung nodules rectal cancer p CCRT[z] VATS[aa] mediastinal tumor excision LAR[ab]
0.042	0	right ACL[ac] MCL[ad] injury arthroscopy ACL reconstruction
0.041	0	1 C4 5 6 spondylosis 2 right carpal tunnel 1 ACDF[ae] C4 5 6 2 right median nerve decompression
0.031	0	bil[af] ov[ag] teratoma laparoscopy adnexectomy
0.030	0	left ureter stone URSL[ah] laser left
0.029	0	uterine myoma robotic myomectomy
0.029	0	acute appendicitis laparoscopic appendectomy
0.029	0	hemorrhoids hemorrhoidectomy
0.029	0	nontoxic goiter thyroidectomy
0.027	0	infertility TVOR[ai]
0.027	0	endometrial polyp TCR[aj]
0.027	0	GA[ak] 38 weeks breech caesarean section
0.025	0	rt[al] breast lesion MRI[am] guided biopsy
0.025	0	right inguinal hernia TEP[an] right

[a]IHCA: intrahospital cardiac arrest.

[b]p: post.

[c]CPR: cardiopulmonary resuscitation.

[d]ECMO: extracorporeal membrane oxygenation.

[e]ACS: acute coronary syndrome.

[f]AR: aortic regurgitation.
[g]CABG: coronary artery bypass graft.
[h]AVR: aortic valve replacement.
[i]AMI: acute myocardial infarction.
[j]TEE: transesophageal echocardiography.
[k]r: rule out.
[l]PPU: perforated peptic ulcer.
[m]LPPU: laparoscopic perforated peptic ulcer surgery.
[n]EXP LAP: exploratory laparotomy.
[o]cath: catheter.
[p]UPPP: uvulopalatopharyngoplasty.
[q]ESRD, end-stage renal disease.
[r]HD: hemodialysis.
[s]perm cath: permanent catheter.
[t]AVG: arteriovenous graft.
[u]AVF: arteriovenous fistula.
[v]PTA: percutaneous transluminal angioplasty.
[w]RLL: right lower lobe.
[x]NF: necrotizing fasciitis.
[y]BK: below-knee amputation.
[z]CCRT: concurrent chemoradiotherapy.
[aa]VATS: video-assisted thoracic surgery.
[ab]LAR: low anterior resection.
[ac]ACL: anterior cruciate ligament.
[ad]MCL: medial collateral ligament.
[ae]ACDF: anterior cervical discectomy and fusion.
[af]bil: bilateral.
[ag]ov: ovarian.
[ah]URSL: ureteroscopic lithotomy.
[ai]TVOR: transvaginal oocyte retrieval.
[aj]TCR: transcervical resectoscope.
[ak]GA: gestational age.
[al]rt: right.
[am]MRI: magnetic resonance imaging.
[an]TEP: total extraperitoneal approach.

Discussion

Principal Findings

The DNN-BERT model predicted the in-hospital 30-day mortality with the highest AUROC of 0.964 (95% CI 0.961-0.967) and an AUPRC of 0.336 (95% CI 0.276-0.402); see Table 3 and Figure 3. The BERT-DNN had an AUROC significantly higher compared to XGBoost, logistic regression, and ASAPS but not the DNN or random forest. The BERT-DNN also had an AUPRC significantly higher compared to the DNN, random forest, XGBoost, logistic regression, and ASAPS.

Hill et al [6] proposed an ML model that outperformed previous tools (eg, preoperative score to predict postoperative mortality, Charlson comorbidity, and ASAPS) and could be used independently by clinicians. Our BERT-DNN model outperformed Hill et al's [6] model, obtaining a higher AUROC, sensitivity, and F1 score than their results (0.964, 95% CI 0.961-0.967 vs 0.932, 95% CI 0.910-0.951; 0.650, 95% CI 0.587-0.719 vs 0.239, 95% CI 0.127-0.379; and 0.347, 95% CI 0.305-0.388 vs 0.302, 95% CI 0.172-0.449, respectively); see Table 3. The preoperative diagnosis text features and proposed procedure information might contribute to our BERT-DNN model and enhance its sensitivity and F1 score. Unlike Hill et al [6], who focused on patients undergoing general anesthesia, we trained and tested our model on both general and neuraxial anesthesia. The DL model with clinical text predicted postoperative mortality significantly more discriminatively than logistic regression and ASAPS (Table 4).

DL methods predict postoperative mortality using preoperative and intraoperative features [7-9]. Using a summary of intraoperative features alongside the ASAPS, Lee et al [7] presented a DNN model that achieved an AUROC of 0.91 (95% CI 0.88-0.93). Our DNN model obtained a higher AUROC than their model because we included key features such as preoperative location and surgical department, the importance of which was also verified in previous studies [6]. Fritz et al [8] proposed a multipath convolutional neural network model to predict postoperative mortality using intraoperative time-series data and preoperative features. Their model achieved an

AUROC of 0.910 (95% CI 0.897-0.924) and an AUPRC of 0.325 (95% CI 0.280-0.372) [33]. In contrast, our model can be used preoperatively and achieve a higher AUROC and AUPRC (Table 3).

Previous studies used *ICD* and CPT codes as categorical features to stratify surgery risk [2,6,7,9,12]. This input feature has many classes, which resulted in a sparse input matrix; this made it difficult for the model to learn helpful information. However, because *ICD* codes were typically recorded after surgery, including them in the preoperative model was impractical. Furthermore, the CPT code was not used globally. For this reason, we could not compare a model including word embeddings with one including CPT codes. However, our results exhibited excellent discrimination with a high AUROC and AUPRC. The AUPRC is significantly higher than models without text. The calibration plot also strongly correlated the predicted probabilities and observed mortalities (Figure 4). Word embedding visualizations showed that the increased predicted probabilities were concordant with high-risk surgery and an increased mortality rate (Figure 5 and Table 6). We showed that word embeddings for surgery information could be used in DL models to predict postoperative mortality before surgery without requiring CPT or *ICD* codes.

The fusion of neural networks, combining diverse types of data (eg, image [34] and time-series [8] data) with 1D data (eg, categorical, and continuous data), improved the model's performance. Including unstructured clinical text via natural language procession can improve intensive care unit (ICU) mortality predictions [14,16]. The DL model that combined unstructured and structured data outperformed models using either type of data alone [15]. Moreover, the performance of the clinical pretrained DL language model could be maintained between different institutions [35].

Limitations

Our study has several limitations. First, postoperative mortality accounted for 1.3% (1562/121,313) of our cohort, and the classes were highly imbalanced. The model training and performance metric evaluations were difficult to apply with these sparse positive labels. To compensate for the class imbalance via an algorithmic method, we applied cost-sensitive learning by balancing the weights of the loss function to emphasize the minority group [36]. We evaluated the discrimination of our model with the AUPRC, which is more informative than the AUROC for imbalanced data [8,25,26]. Second, our model predicted mortality using EHRs. The errors in the records and missing values affected the prediction results. Typos of text interfered with the word-embedding process. Outliers were detected and input using the defined rules (Multimedia Appendix 2). Third, all records were collected from a single large medical center. Although the pipeline we created ensured that the DL model could be reproduced in other institutes, the model weights might vary for a different data set. The generalizability of our results must be examined in future studies.

Conclusion

In conclusion, descriptive surgical text was essential for predicting postoperative mortality. The word embeddings of preoperative diagnoses and proposed procedures, via the contextualized language model BERT, were combined in DL models to predict postoperative mortality. This predictive capacity can help identify patients with higher risk from structure data and text of EHRs.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan (Grant MOST 110-2634-F-002-032-), and the Far Eastern Memorial Hospital, Taiwan (Grant FEMH-2021-C-056). The sponsors had no role in the study design, data collection and analysis, publication decision, or manuscript drafting. All authors have approved the final article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of laboratory values and vital sign values.
[DOC File , 45 KB - medinform_v10i5e38241_app1.doc]

Multimedia Appendix 2
Continuous feature limits to define outliers.
[DOC File , 33 KB - medinform_v10i5e38241_app2.doc]

Multimedia Appendix 3
Higher resolution of Figure 5. Word embeddings visualized by t distributed stochastic neighbor embedding. (A) Word embeddings of the training set. (B) Word embeddings of the testing set. "Probs" indicates probabilities predicated by the BERT-DNN model. The intensity of color increased with the probability. "Labels" indicates mortalities by "x" and survivors by "•". ards: acute respiratory distress syndrome; atfl: anterior talofibular ligament; avg: arteriovenous graft; avp: aortic valvuloplasty; BERT: bidirectional encoder representations from transformers; bct: breast-conserving therapy; bil: bilateral; bph: benign prostate

hypertrophy; bx: biopsy; chr: chronic hypertrophic rhinitis; cps: chronic paranasal sinusitis; dbj: double J stent; DNN: deep neural network; ecmo: extracorporeal membrane oxygenation; emh: endometrial hemorrhage; esrd: end-stage renal disease; fess: functional endoscopic sinus surgery; itc: intertrochanter; ivg: intravenous general anesthesia; lih: left inguinal hernia; mvr: mitral valve replacement; nsd: nasal septum deviation; p: post; pcnl: percutaneous nephrolithotomy; perm cath: permanent catheter; psa: prostate-specific antigen; r: rule out; r't: right; rirs: retrograde intrarenal surgery; rv: right ventricle; slnd: sentinel lymph node dissection; SNE: stochastic neighbor embedding; t colon: transverse colon; tee: transesophageal echocardiography; tep: total extraperitoneal approach; trus: transrectal ultrasound; turp: transurethral resection of the prostate; urs: ureteroscopy; vats: video-assisted thoracic surgery; vhd: valvular heart disease.
[PNG File , 5102 KB - medinform_v10i5e38241_app3.png]

References

1. International Surgical Outcomes Study Group. Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries. Br J Anaesth 2016 Oct 31;117(5):601-609 [FREE Full text] [doi: 10.1093/bja/aew316] [Medline: 27799174]
2. Sigakis M, Bittner E, Wanderer J. Validation of a risk stratification index and risk quantification index for predicting patient outcomes: in-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. Anesthesiology 2013 Sep;119(3):525-540 [FREE Full text] [doi: 10.1097/ALN.0b013e31829ce6e6] [Medline: 23770598]
3. Le Manach Y, Collins G, Rodseth R, Le Bihan-Benjamin C, Biccard B, Riou B, et al. Preoperative Score to Predict Postoperative Mortality (POSPOM): derivation and validation. Anesthesiology 2016 Mar;124(3):570-579 [FREE Full text] [doi: 10.1097/ALN.0000000000000972] [Medline: 26655494]
4. Mayhew D, Mendonca V, Murthy BVS. A review of ASA physical status: historical perspectives and modern developments. Anaesthesia 2019 Mar 15;74(3):373-379 [FREE Full text] [doi: 10.1111/anae.14569] [Medline: 30648259]
5. Bilimoria K, Liu Y, Paruch J, Zhou L, Kmiecik T, Ko C, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013 Nov;217(5):833-42.e1 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2013.07.385] [Medline: 24055383]
6. Hill BL, Brown R, Gabel E, Rakocz N, Lee C, Cannesson M, et al. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. Br J Anaesth 2019 Dec;123(6):877-886 [FREE Full text] [doi: 10.1016/j.bja.2019.07.030] [Medline: 31627890]
7. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. Anesthesiology 2018 Oct;129(4):649-662 [FREE Full text] [doi: 10.1097/ALN.0000000000002186] [Medline: 29664888]
8. Fritz BA, Cui Z, Zhang M, He Y, Chen Y, Kronzer A, et al. Deep-learning model for predicting 30-day postoperative mortality. Br J Anaesth 2019 Nov;123(5):688-695 [FREE Full text] [doi: 10.1016/j.bja.2019.07.025] [Medline: 31558311]
9. Yan X, Goldsmith J, Mohan S, Turnbull Z, Freundlich R, Billings F, et al. Impact of intraoperative data on risk prediction for mortality after intra-abdominal surgery. Anesth Analg 2022 Jan 01;134(1):102-113. [doi: 10.1213/ANE.0000000000005694] [Medline: 34908548]
10. Konishi T, Goto T, Fujiogi M, Michihata N, Kumazawa R, Matsui H, et al. New machine learning scoring system for predicting postoperative mortality in gastroduodenal ulcer perforation: a study using a Japanese nationwide inpatient database. Surgery 2022 Apr;171(4):1036-1042. [doi: 10.1016/j.surg.2021.08.031] [Medline: 34538648]
11. Rogers MP, Janjua H, DeSantis AJ, Grimsley E, Pietrobon R, Kuo PC. Machine learning refinement of the NSQIP risk calculator: who survives the "Hail Mary" case? J Am Coll Surg 2022 Apr 01;234(4):652-659. [doi: 10.1097/XCS.0000000000000108] [Medline: 35290285]
12. Dalton J, Kurz A, Turan A, Mascha E, Sessler D, Saager L. Development and validation of a risk quantification index for 30-day postoperative mortality and morbidity in noncardiac surgical patients. Anesthesiology 2011 Jun;114(6):1336-1344 [FREE Full text] [doi: 10.1097/ALN.0b013e318219d5f9] [Medline: 21519230]
13. Hashimoto D, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. Anesthesiology 2020 Feb;132(2):379-394 [FREE Full text] [doi: 10.1097/ALN.0000000000002960] [Medline: 31939856]
14. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, Himes BE, et al. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. Crit Care Med 2018 Jul;46(7):1125-1132 [FREE Full text] [doi: 10.1097/CCM.0000000000003148] [Medline: 29629986]
15. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 2020 Oct 29;20(1):280 [FREE Full text] [doi: 10.1186/s12911-020-01297-6] [Medline: 33121479]
16. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. JAMA Netw Open 2018 Dec 07;1(8):e185097 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.5097] [Medline: 30646310]
17. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018 Oct 11. [doi: 10.48550/arXiv.1810.04805]

18. Arnaud, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. 2022 Presented at: Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - Scale-IT-up; 2022; Vienna p. 835. [doi: 10.5220/0011012800003123]
19. Kades K, Sellner J, Koehler G, Full PM, Lai TYE, Kleesiek J, et al. Adapting bidirectional encoder representations from transformers (BERT) to assess clinical semantic textual similarity: algorithm development and validation study. JMIR Med Inform 2021 Feb 03;9(2):e22795 [FREE Full text] [doi: 10.2196/22795] [Medline: 33533728]
20. Loper E, Bird S. Nltk: The natural language toolkit. arXiv 2002 May 17. [doi: 10.48550/arXiv.cs/0205028]
21. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
22. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. 2016 Aug Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
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. Alsentzer E, Murphy J, Boag W, Weng W, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv 2019 Apr 06. [doi: 10.18653/v1/w19-1909]
25. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
26. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015 Aug;68(8):855-859 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.02.010] [Medline: 25881487]
27. DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837. [doi: 10.2307/2531595]
28. Boyd K, Eng K, Page C. Area under the precision-recall curve: point estimates and confidence intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, editors. Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer; 2013:451-466.
29. Maaten LVD, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605.
30. Jin M, Bahadori M, Colak A, Bhatia P, Celikkaya B, Bhakta R, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. arXiv 2018 Nov 29.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Kossaifi J, Thirion B, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011 Feb 01;12:2825-2830.
32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. 2019 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS); Dec 8-14, 2019; Vancouver, Canada p. A.
33. Fritz BA, Abdelhack M, King CR, Chen Y, Avidan MS. Update to 'Deep-learning model for predicting 30-day postoperative mortality' (Br J Anaesth 2019; 123: 688-95). Br J Anaesth 2020 Aug;125(2):e230-e231 [FREE Full text] [doi: 10.1016/j.bja.2020.04.010] [Medline: 32389391]
34. Huang S, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 2020 Oct 16;3(1):136 [FREE Full text] [doi: 10.1038/s41746-020-00341-z] [Medline: 33083571]
35. Bear Don't Walk Iv OJ, Sun T, Perotte A, Elhadad N. Clinically relevant pretraining is all you need. J Am Med Inform Assoc 2021 Aug 13;28(9):1970-1976. [doi: 10.1093/jamia/ocab086] [Medline: 34151966]
36. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019 Mar 19;6(1):1-54. [doi: 10.1186/s40537-019-0192-5]

Abbreviations

ASAPS: American Society of Anesthesiologist Physical Status
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operator characteristic curve
BERT: bidirectional encoder representations from transformers
CPT: Current Procedural Terminology
DL: deep learning
DNN: deep neural network
EHR: electronic health record
FC: fully connected
ICD: International Classification of Diseases
ML: machine learning
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 24.03.22; peer-reviewed by M Elbattah; comments to author 14.04.22; revised version received 18.04.22; accepted 26.04.22; published 10.05.22.

Please cite as:
Chen PF, Chen L, Lin YK, Li GH, Lai F, Lu CW, Yang CY, Chen KC, Lin TY
Predicting Postoperative Mortality With Deep Neural Networks and Natural Language Processing: Model Development and Validation
JMIR Med Inform 2022;10(5):e38241
URL: https://medinform.jmir.org/2022/5/e38241
doi:10.2196/38241
PMID:35536634

Original Paper

Electronic Medical Record–Based Machine Learning Approach to Predict the Risk of 30-Day Adverse Cardiac Events After Invasive Coronary Treatment: Machine Learning Model Development and Validation

Osung Kwon[1*], MD, PhD; Wonjun Na[2*], MS; Heejun Kang[3], MS; Tae Joon Jun[3], PhD; Jihoon Kweon[3], PhD; Gyung-Min Park[4], MD, PhD; YongHyun Cho[5], MS; Cinyoung Hur[5], MS; Jungwoo Chae[5], BS; Do-Yoon Kang[3], MD, PhD; Pil Hyung Lee[3], MD, PhD; Jung-Min Ahn[3], MD, PhD; Duk-Woo Park[3], MD, PhD; Soo-Jin Kang[3], MD, PhD; Seung-Whan Lee[3], MD, PhD; Cheol Whan Lee[3], MD, PhD; Seong-Wook Park[3], MD, PhD; Seung-Jung Park[3], MD, PhD; Dong Hyun Yang[6*], MD, PhD; Young-Hak Kim[3*], MD, PhD

[1]Division of Cardiology Department of Internal Medicine, Eunpyeong St Mary's Hospital, Catholic University of Korea, Seoul, Republic of Korea

[2]Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[3]Division of Cardiology, Department of Internal Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[4]Division of Cardiology, Department of Internal Medicine, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan, Republic of Korea

[5]Artificial Intelligence Lab, Linewalks, Inc, Seoul, Republic of Korea

[6]Department of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Young-Hak Kim, MD, PhD
Division of Cardiology, Department of Internal Medicine, Asan Medical Center
University of Ulsan College of Medicine
88 Olympic-ro 43-gil, Songpa-gu
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 3995
Fax: 82 2 486 5918
Email: mdyhkim@amc.seoul.kr

Abstract

Background: Although there is a growing interest in prediction models based on electronic medical records (EMRs) to identify patients at risk of adverse cardiac events following invasive coronary treatment, robust models fully utilizing EMR data are limited.

Objective: We aimed to develop and validate machine learning (ML) models by using diverse fields of EMR to predict the risk of 30-day adverse cardiac events after percutaneous intervention or bypass surgery.

Methods: EMR data of 5,184,565 records of 16,793 patients at a quaternary hospital between 2006 and 2016 were categorized into static basic (eg, demographics), dynamic time-series (eg, laboratory values), and cardiac-specific data (eg, coronary angiography). The data were randomly split into training, tuning, and testing sets in a ratio of 3:1:1. Each model was evaluated with 5-fold cross-validation and with an external EMR-based cohort at a tertiary hospital. Logistic regression (LR), random forest (RF), gradient boosting machine (GBM), and feedforward neural network (FNN) algorithms were applied. The primary outcome was 30-day mortality following invasive treatment.

Results: GBM showed the best performance with area under the receiver operating characteristic curve (AUROC) of 0.99; RF had a similar AUROC of 0.98. AUROCs of FNN and LR were 0.96 and 0.93, respectively. GBM had the highest area under the precision-recall curve (AUPRC) of 0.80, and the AUPRCs of RF, LR, and FNN were 0.73, 0.68, and 0.63, respectively. All models showed low Brier scores of <0.1 as well as highly fitted calibration plots, indicating a good fit of the ML-based models.

On external validation, the GBM model demonstrated maximal performance with an AUROC of 0.90, while FNN had an AUROC of 0.85. The AUROCs of LR and RF were slightly lower at 0.80 and 0.79, respectively. The AUPRCs of GBM, LR, and FNN were similar at 0.47, 0.43, and 0.41, respectively, while that of RF was lower at 0.33. Among the categories in the GBM model, time-series dynamic data demonstrated a high AUROC of >0.95, contributing majorly to the excellent results.

Conclusions: Exploiting the diverse fields of the EMR data set, the ML-based 30-day adverse cardiac event prediction models demonstrated outstanding results, and the applied framework could be generalized for various health care prediction models.

(JMIR Med Inform 2022;10(5):e26801) doi:10.2196/26801

KEYWORDS

big data; electronic medical record; machine learning; mortality; adverse cardiac event; coronary artery disease; prediction

Introduction

Cardiovascular disease is the leading cause of mortality throughout the world and is associated with various morbidities [1]. Invasive treatment, including percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG) surgery, is commonly required in patients with acute coronary syndrome and stable angina. Owing to the potential risk associated with inevitable invasiveness and the individual comorbidities, risk stratification and identification of high-risk patients is warranted [2,3]. Accordingly, several risk prediction models for adverse events after invasive coronary treatment have been proposed [4-7]. However, their use is limited owing to inadequate predictive ability, low generalizability, and lack of individualized risk assessment, as they have been developed using limited number of variables in select cohorts.

In recent times, with an increase in the availability of large volume of electronic medical record (EMR) data, there has been a gradual interest in using data-driven approaches to construct efficient tools for risk prediction [8,9]. In addition, machine learning (ML) algorithms are gaining popularity as an alternative approach for risk prediction to deal with complex EMR data and to overcome the limitations of previous models [10]. Recent work on models based on EMR data for predicting adverse events suggests that incorporation of ML might allow more accurate risk prediction [11-14]. However, validated robust models are still limited, as the previous models used prespecified variables based on traditional risk factors mainly comprising structural data or lacked proper external validation. Thus, this study aimed to develop ML models by utilizing diverse fields of both structured and unstructured EMR data to predict the risk of 30-day major adverse cardiac events (MACE), including mortality, after PCI or CABG and to validate the model in a different cohort.

Methods

Database

Development and Internal Validation Set

The data for this study were obtained from Asan Medical Center, which provides quaternary medical care for people in South Korea. It has 55 departments—approximately 2700 beds—and >8000 employees; it sees approximately 3,000,000 outpatient clinic visits and 900,000 admissions per year. The Asan biomedical research environment is the data warehouse system of Asan Medical Center, which has deidentified information of 4 million patients and is updated every 3 days [15]. The Asan heart registry was constructed from diverse fields of structured or unstructured EMR data extracted from the Asan biomedical research environment database by using structured query language. The registry comprised 571,157 patients, and the inclusion criteria were inpatient admissions or outpatient visits in the cardiology, cardiac surgery, or emergency department for established or suspected heart diseases between January 1, 2000 and November 30, 2016.

External Validation Set

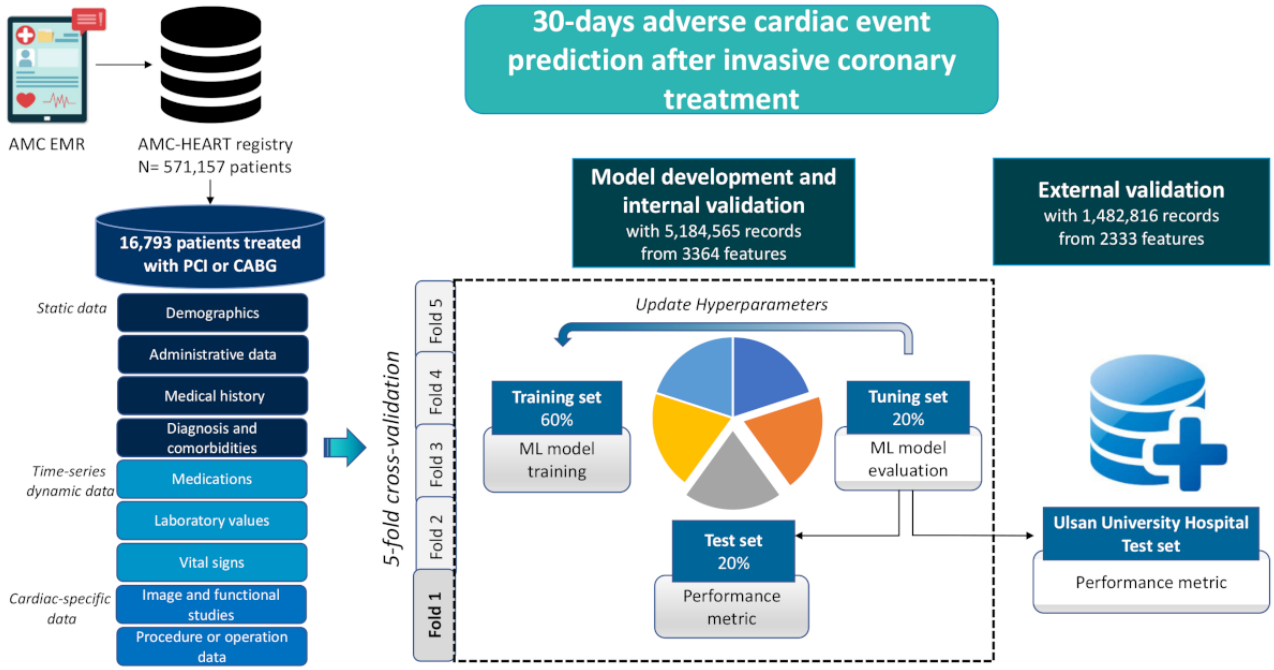
For external validation, we used data obtained from the EMRs of Ulsan University Hospital, which is a tertiary hospital with approximately 900 beds that caters to a metropolitan city and its surrounding suburban area in the southern region of South Korea. The patients' demographics, medical practice, and operating systems differ between the 2 hospitals, which would allow evaluation of the model in a different population.

Data Processing

The overall process for building the EMR-based database is presented in Figure S1 of Multimedia Appendix 1. Briefly, first, we collected the anonymized records of 748,474 patients who had visited the Asan Medical Center or Ulsan University Hospital because of cardiovascular diseases. Second, we set clinically plausible criteria to remove errors and duplications. Third, we integrated unstructured data such as readings of medical examinations with structured data sourced from EMRs to create the CardioNet [16]. We subsequently performed text mining to structuralize the significant variables associated with cardiovascular diseases because most results of the principal cardiovascular diseases–related medical examinations are free-text readings. The basic method of text mining applied to unstructured data can be described in 3 steps. First, we created a metatable consisting of the main variables and conditions of extraction by the clinician. Second, we divided the readings into 3 frames: text, tabular, and others, and defined the extraction rules for each frame. We took into consideration the structure of the original data and the location of variables set in the metatable and defined rules by using a variety of operators and regular expressions. Third, the new tables were built by extracting the keywords and features from the original data. The values of the keywords were based on rules defined in the previous step. Additionally, to ensure interoperability for convergent multicenter research, we standardized the data by using several codes that correspond to the common data model.

Finally, we created the descriptive table (ie, dictionary of the CardioNet) to simplify access and utilization of data for clinicians and engineers and continuously validated the data to ensure reliability [16]. Most structured data were obtained using classic preprocessing technologies, including data cleansing, data integration, data transformation, data reduction, and privacy protection. Finally, we extracted the following structured data elements: demographics, administrative information, medical history and comorbidities, diagnoses, vital signs, laboratory values, and medications. Unstructured data included the following elements: reports of cardiac-specific studies such as thallium-201 single-photon emission computed tomography (SPECT), coronary angiography, and physicians' procedure notes for PCI or CABG. In this study, we found that with the algorithms developed, we classified the data into 3 categories: basic static data (demographics, administration data, medical history, comorbidities, and diagnosis), dynamic time-series data (medications, laboratory values, and vital signs), and disease-specific data (electrocardiography, treadmill test, echocardiography, coronary computerized tomography, thallium-201 SPECT, coronary angiography, PCI, and CABG) (see Figure 1). The details of the variables in each category are presented in Table S1 in Multimedia Appendix 1 [16]. With respect to the data of the procedures or operation, the variables only confined to the index PCI or CABG were used for this investigation. Data collection and preparation were approved by the Asan Medical Center and Ulsan University Hospital institutional review board, and the requirement for informed consent was waived. Patient deidentification was performed in line with the Health Insurance Portability and Accountability Act. This report adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis reporting guideline [17].

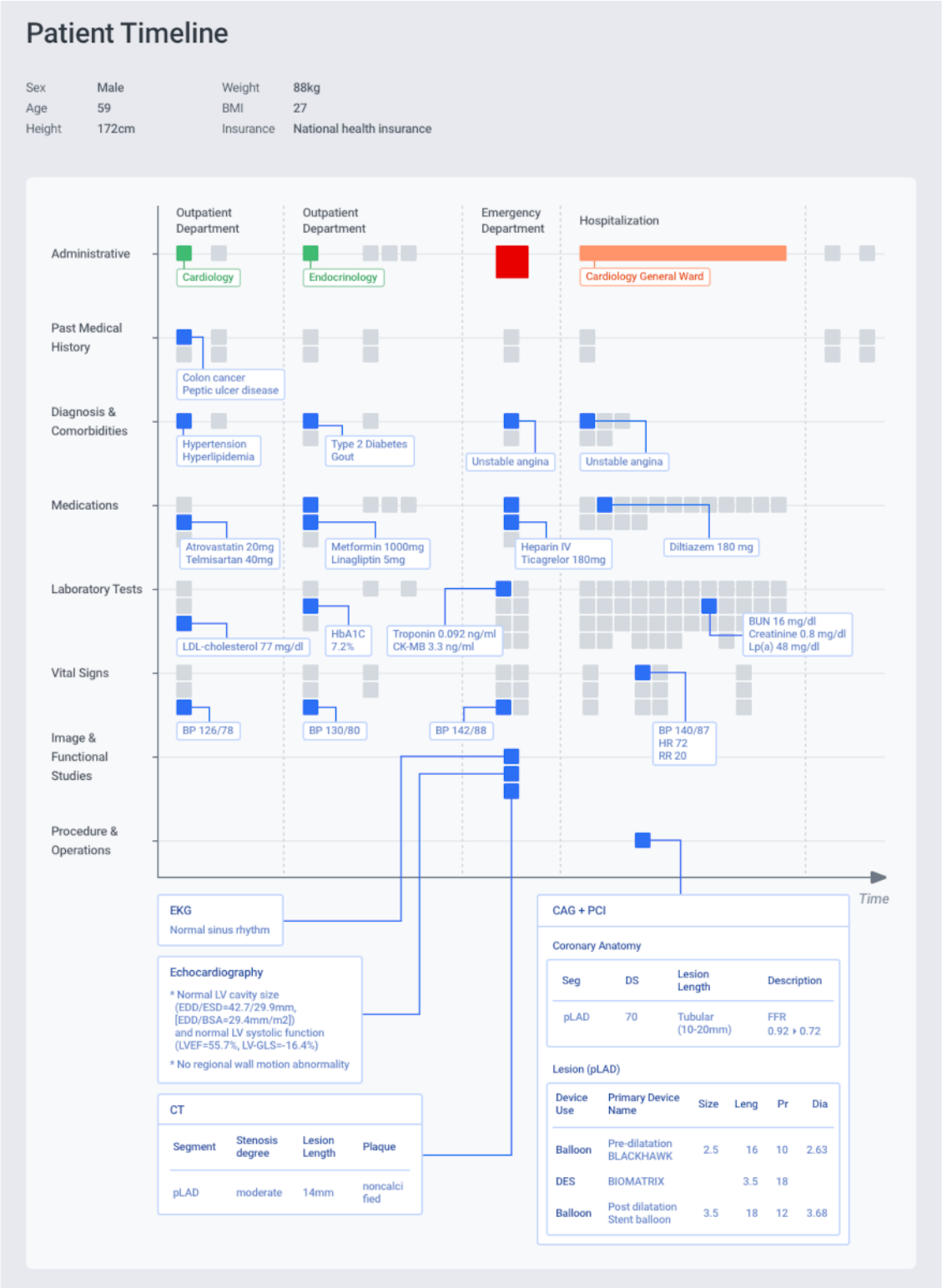
Figure 1. Study diagram. Database, machine learning, and validation. AMC: Asan Medical Center; CABG: coronary artery bypass grafting; EMR: electronic medical record; ML: machine learning; PCI: percutaneous coronary intervention.



Study Population and Outcome

A cohort of 16,793 patients who had undergone PCI (n=12,519) or CABG (n=4274) between January 1, 2006 and November 30, 2016 was identified in the Asan heart registry. As the majority of patients underwent the index PCI or CABG within 1 year after their first generation of data in EMR, we fairly used 1-year accumulated data prior to index procedures for the entire population. The total number of independent records in the data set was 5,184,565, derived from 3364 features. Figure 2 illustrates an example of the patients treated with PCI, encompassing the serial and various EMR data. In the external validation cohort from Ulsan University Hospital, 4159 patients comprising 3950 who underwent PCI and 209 who underwent CABG between January 1, 2006 and November 30, 2016 were included. The data set consisted of 1,482,816 records from 2333 features. Mortality was the primary endpoint, captured through documentation of mortality in the EMR based upon National Health Insurance information. MACE as the secondary endpoint referred to a composite of all-cause mortality, including myocardial infarction, stroke, or repeat revascularization at 30 days following the index invasive treatment. Myocardial infarction, stroke, and repeat revascularization were initially identified from source documents, including diagnosis, electrocardiography, laboratory tests, procedural notes, and results of imaging studies such as magnetic resonance imaging or computerized tomography. Subsequently, the events were rigorously adjudicated by cardiologists or neurologists according to the current definitions [18].

Figure 2. An example case incorporating serial and various electronic medical record data to predict adverse events. BP: blood pressure; BSA: body surface area; BUN: blood urea nitrogen; CAG: coronary angiography; CK-MB: creatine kinase myocardial band; Dia: diameter; EDD: end diastolic dimension; EF: ejection fraction; EKG: electrocardiogram; ESD: end systolic dimension; FFR: fractional flow rate; GLS: global longitudinal strain; Hb: hemoglobin; HR: heart rate; LDL: low-density lipoprotein; Leng: length; Lp(a): lipoprotein A; LV: left ventricle; PCI: percutaneous coronary intervention; pLAD: proximal left anterior descending; Pr: pressure; RR: respiratory rate.



ML Algorithms and Statistics

We only used data generated until index PCI or CABG, whereas data obtained after the index procedure were excluded for developing ML algorithms (see Figure 2). Three approaches were applied to preprocess data generated until index procedures: (1) history-aware encoding is used to reflect whether clinical events had occurred before a certain period of time, (2) one-hot encoding is used to express the existence and missingness of variables, and (3) characteristics of time-series variables were captured by using descriptive statistics (eg, minimum, maximum, average, and count). The detailed

explanation regarding time-series data analysis is shown in Multimedia Appendix 2. The study population was randomly split into training, tuning, and validation cohorts in a ratio of 3:1:1. Four commonly used classes of ML algorithms were used: logistic regression (LR), random forest (RF), gradient boosting machine (GBM), and feedforward neural network (FNN). LR transforms output by using a logistic sigmoid function. RF is an extension of the bagging method, as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees and decide output by majority voting using multiple decision trees. GBM is similar to RF, except that they build 1 tree at a time and combine the voting results in a gradual, additive, and sequential manner. FNN is a classic type of deep learning model that uses hierarchical layers of abstraction and computes the output by using a combination of multiple nodes with nonlinear activation.

The hyperparameters for each model were determined using an empirical search and 5-fold cross-validation on the study population to determine the values that had the best performance (see Figure 1). Hyperparameters and their values in each model are summarized in Table 1. The optimal values of the tuning parameters were identified based on the testing accuracy values that were calculated for each fold and averaged. External validation of the developed prediction models was performed in a cohort from a different hospital. In addition, we determined the performance of each data category and checked the cumulative performance with combinations of multiple information categories, adding each category one by one to identify the best performance. Development of risk algorithms in the training cohort and application of the risk algorithms to the validation cohort was completed using Python with library packages “Keras with Tensorflow backend.” To investigate the important variables in each developed model, we used the permutation feature importance algorithm for LR and FNN, Gini impurity for RF, and frequency of variables for GBM.

Table 1. Hyperparameters and those values of each model.

Model, hyperparameter	Value
Logistic regression	
Solver	liblinear
Maximal iteration	100
Random forest	
Number of estimators	100
Maximal depth	10
Gradient boosting machine	
Objective	binary
Estimators	150
Boosting type	Gradient boosting decision tree
Number of leaves	15
Maximal depth	–1 (no limit)
Learning rate	0.025
Minimal number of data in child	90
Feedforward neural network	
Learning rate	0.0002
Hidden layer units	(64,64)
Batch size	64
Epoch	40
Dropout rate	0.5
Optimizer	Adam (beta1=.5, beta2=.999)

The descriptive characteristics of the study population are provided as number (%) and mean (SD) for categorical and continuous variables, respectively. The discrimination performance of each model was evaluated based on the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). In addition, we evaluated model calibration (ie, the model’s ability to accurately predict the observed absolute risk) by using the Brier score, where 0 would indicate perfect calibration, and generated the calibration plots. A 2-sided P value <.05 was considered indicative of statistical significance. We did not perform any imputation of the missing numerical values, as explicit imputation of missing values does not always provide consistent improvements in predictive models based on electronic health records [19,20]. Because of inevitable differences in the

characteristics and amounts of data between cohorts, we used binary indicators of missingness on external validation [21].

Results

Baseline Characteristics and Event Rates

The baseline characteristics of the population in the development and internal validation groups are listed in Table 2. The mean patient age was 62.7 (SD 10.2) years; of the 16,793 patients, 12,465 (74.2%) were males and 6084 (36.2%) had diabetes, while 243 (1.4%) had a history of congestive heart failure. Chronic renal insufficiency, chronic lung disease, and chronic liver disease were reported in 566 (3.4%), 386 (2.3%), and 487 (2.9%) patients, respectively. Approximately two-thirds of the patients were admitted via outpatient clinics while the remaining patients were admitted via the emergency department. Among 16,793 patients in our developmental cohort, MACE at 30 days occurred in 1500 (8.9%) patients, including 178 cases (1.1%) of mortality, 1159 (6.9%) cases of myocardial infarction, 124 (7.4%) cases of stroke, and 180 (1.1%) cases of repeat revascularization. Among a total of 4159 patients in the external validation cohort, there were 75 (1.8%) mortalities at 30 days follow-up; the details of the patients' characteristics in the external validation cohort are shown in Table 3.

Table 2. Baseline clinical characteristics of the development and internal validation set.

Characteristics	Development and internal validation set		
	Total population (N=16,793)	Percutaneous coronary intervention (n=12,519)	Coronary artery bypass grafting surgery (n=4274)
Age (years), mean (SD)	62.7 (10.2)	62.2 (10.5)	64.1 (9.4)
Male sex, n (%)	12,465 (74.2)	9312 (74.4)	3153 (73.8)
Body mass index (kg/m^2), mean (SD)	24.9 (3.1)	25.0 (3.0)	24.6 (3.1)
Hypertension, n (%)	10,697 (63.7)	7758 (62)	2939 (68.8)
Diabetes mellitus, n (%)	6084 (36.2)	4127 (33)	1957 (45.8)
Hyperlipidemia, n (%)	9200 (54.8)	6932 (55.4)	2268 (53.1)
Current cigarette smoker, n (%)	3009 (17.9)	2424 (19.4)	585 (13.7)
Prior myocardial infarction, n (%)	568 (3.4)	394 (3.1)	174 (4.1)
Previous cerebrovascular accident, n (%)	596 (3.5)	420 (3.4)	176 (4.1)
History of congestive heart failure, n (%)	243 (1.4)	132 (1.1)	111 (2.6)
Peripheral vascular disease, n (%)	278 (1.7)	199 (1.6)	79 (1.8)
Valvular heart disease, n (%)	387 (2.3)	106 (0.8)	281 (6.6)
Chronic renal insufficiency, n (%)	566 (3.4)	363 (2.9)	203 (4.7)
Chronic lung disease, n (%)	386 (2.3)	306 (2.4)	80 (1.9)
Chronic liver disease, n (%)	487 (2.9)	396 (3.2)	91 (2.1)
History of malignancy, n (%)	1019 (6.1)	816 (6.5)	203 (4.7)
Presentation with acute myocardial infarction, n (%)	3032 (18.1)	2509 (20)	523 (12.2)
Admission via emergency department, n (%)	5054 (30.1)	3941 (31.5)	1113 (26)
Admission via outpatient clinics, n (%)	11,739 (69.9)	8578 (68.5)	3161 (74)

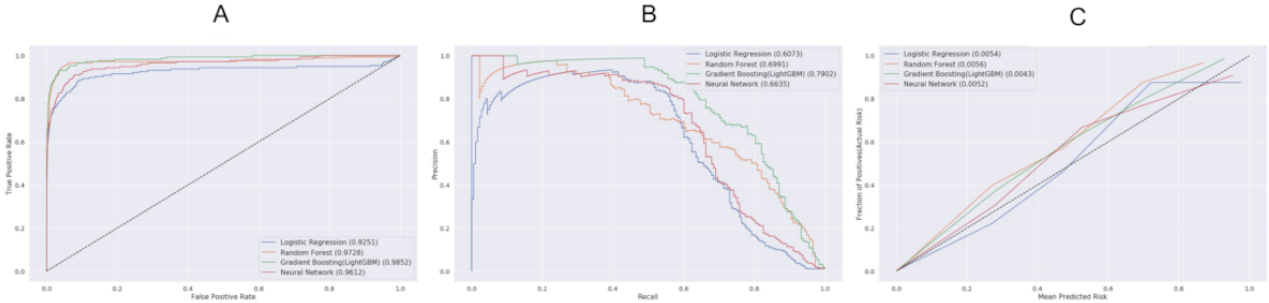
Table 3. Baseline clinical characteristics of the external validation set.

Characteristics	External validation set		
	Total population (n=4159)	Percutaneous coronary intervention (n=3950)	Coronary artery bypass grafting surgery (n=209)
Age (years), mean (SD)	61.7 (10.9)	61.6 (9.4)	62.7 (10.9)
Male sex, n (%)	2913 (70)	2779 (70.3)	134 (64.1)
Body mass index (kg/m^2), mean (SD)	24.0 (5.4)	24.0 (5.2)	23.8 (6.4)
Hypertension, n (%)	1947 (46.8)	1851 (46.8)	96 (45.9)
Diabetes mellitus, n (%)	1278 (30.7)	1195 (30.2)	83 (39.7)
Hyperlipidemia, n (%)	1154 (27.7)	1098 (27.7)	56 (26.7)
Current cigarette smoker, n (%)	1285 (30.9)	1234 (31.2)	51 (24.4)
Prior myocardial infarction, n (%)	280 (6.7)	265 (6.7)	15 (7.1)
Previous cerebrovascular accident, n (%)	233 (5.6)	220 (5.5)	13 (6.2)
History of congestive heart failure, n (%)	76 (1.8)	71 (1.7)	5 (2.3)
Peripheral vascular disease, n (%)	49 (1.1)	45 (1.1)	4 (1.9)
Valvular heart disease, n (%)	27 (0.6)	18 (0.4)	9 (4.3)
Chronic renal insufficiency, n (%)	130 (3.1)	123 (3.1)	7 (3.3)
Chronic lung disease, n (%)	146 (3.5)	143 (3.6)	3 (1.4)
Chronic liver disease, n (%)	201 (4.8)	193 (4.8)	8 (3.8)
History of malignancy, n (%)	192 (4.6)	183 (4.6)	9 (4.3)
Presentation with acute myocardial infarction, n (%)	1357 (32.6)	1314 (33.2)	43 (20.5)
Admission via emergency department, n (%)	1706 (41)	1634 (41.3)	72 (34.4)
Admission via outpatient clinics, n (%)	2453 (58.9)	2316 (58.6)	137 (65.5)

Performance in Predicting 30-Day Mortality

Figure 3 demonstrates the discrimination and calibration results of 5-fold cross-validation obtained by evaluation with each technique. GBM showed the highest AUROC with a value of 0.99 (95% CI 0.97-0.99, *P*<.001) and RF showed similar AUROC of 0.98 (95% CI 0.96-0.0.99, *P*<.001) (see Figure 3A). The AUROCs of FNN and LR were slightly lower at 0.96 (95% CI 0.93-0.99, *P*<.001) and 0.93 (95% CI 0.87-0.99, *P*<.001), respectively. GBM had the highest AUPRC with a value of 0.80, and AUPRCs of RF, LR, and FNN were 0.73, 0.68, and 0.63, respectively (see Figure 3B). In terms of model calibration, all models showed low Brier scores of less than 0.1, indicating an excellent fit of the ML-based models (see Figure 3C). Calibration plots for each model also confirmed good agreement between the estimated predicted risk and observed risk.

Figure 3. Five-fold cross-validation of performance of each machine model in predicting 30-day mortality after invasive treatment. A. Area under the receiver-operator characteristic curve, B. Area under the precision-recall curve, and C. Calibration plot with Brier score.



On external validation using the data set of the Ulsan University hospital, maximal predictive performance was observed with GBM (AUROC 0.90, 95% CI 0.86-0.95; *P*<.001), followed by FNN with AUROC of 0.85 (95% CI 0.81-0.92, *P*<.001) (see Figure 4A). LR and RF showed slightly lower AUROCs of 0.80 (95% CI 0.73-0.87, *P*<.001) and 0.79 (95% CI 0.74-0.84, *P*<.001), respectively. The AUPRCs of GBM, LR, and FNN showed similar values of 0.47, 0.42, and 0.41, respectively; however, that of RF was lower at 0.33 (see Figure 4B). All

models showed low Brier scores of <0.1, indicating a good fit of the ML-based models (see Figure 4C).

Figure 4. External validation of performance of each machine model in predicting 30-day mortality after invasive treatment. A. Area under the receiver operator characteristic curve, B. Area under the precision-recall curve, and C. Calibration plot with Brier score.

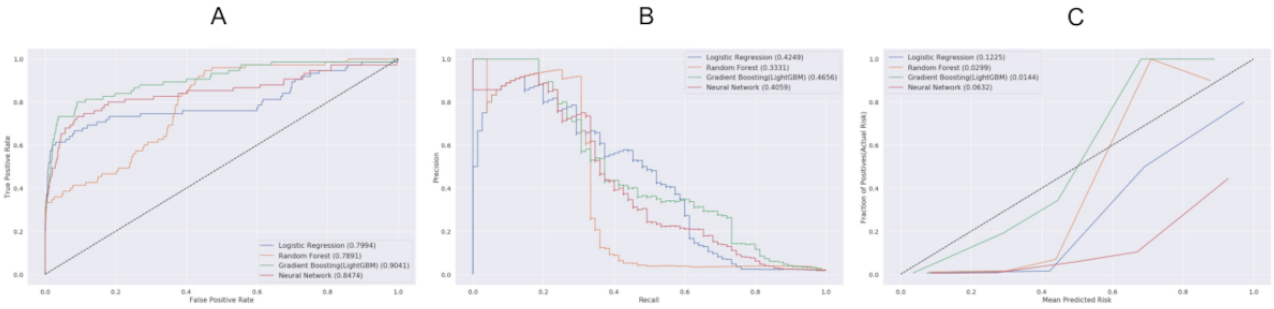
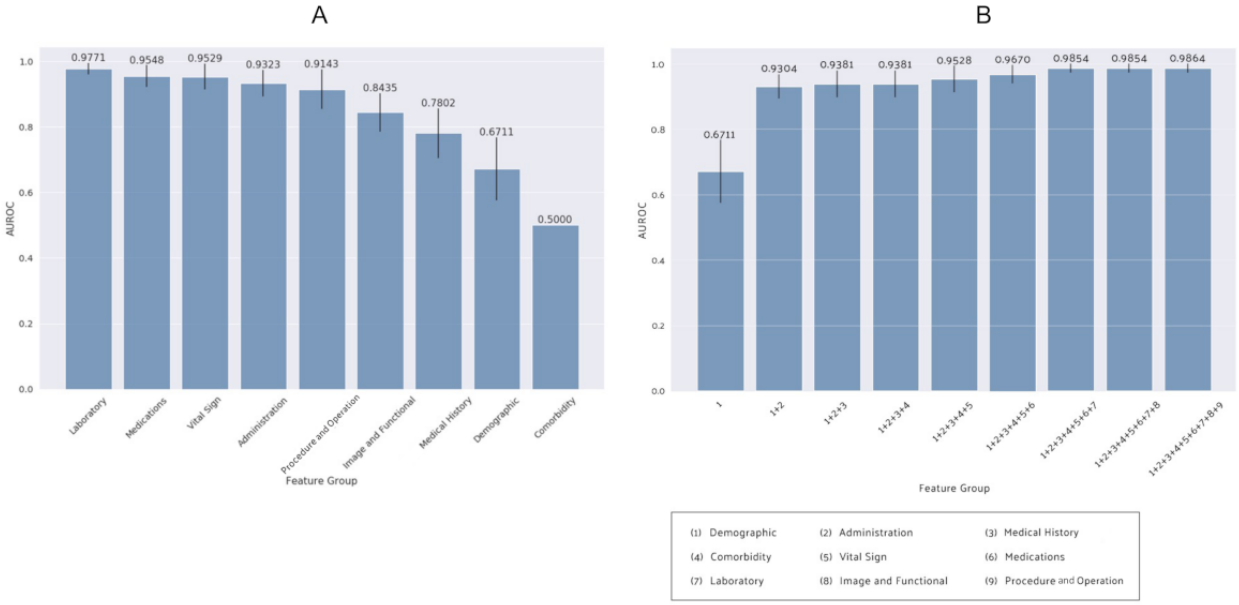


Figure 5A illustrates the predictive performance of each data category in GBM, which showed the highest AUROC. Among the individual categories, laboratory values demonstrated the highest AUROC with a value of 0.98. Medications and vital signs showed the second highest AUROCs with a value of 0.95. In contrast, static data such as diagnosis and comorbidities category, data, and medical history showed low AUROCs of <0.80. GBM using combinations of feature categories showed progressive improvement in performance, while dynamic time-series data was gradually included on top of the basic static data, after which subtle improvement was seen when adding cardiac-specific data (see Figure 5B).

Figure 5. Prediction performance of the gradient boosting machine model assessed by area under the receiver operator characteristic curves. A. Each data category, B. Combination of data categories. AUROC: area under the receiver operator characteristic curve.



Performance in Predicting MACE

The performance of the ML models for predicting 30-day MACE is demonstrated in Table 4. The maximal predictive performance was observed with GBM (AUROC 0.88, 95% CI 0.85-0.90; *P*<.001). RF and FNN showed a similar performance with AUROCs of 0.85 (95% CI 0.83-0.88, *P*<.001) and 0.85 (95% CI 0.83-0.88, *P*<.001), respectively, while the AUROC of the LR was lower at 0.83 (95% CI 0.82-0.88, *P*<.001). In terms of the AUPRC, GBM showed the highest value of 0.50, followed by FNN, RF, and LR with values of 0.41, 0.39, and 0.37, respectively. All models showed low Brier scores of less than 0.1, indicating a good fit of the ML-based models.

Table 4. Performance of machine learning models for predicting major adverse cardiac events.

Model	Area under the receiver operating characteristic curve	95% CI	*P* value	Area under the precision-recall curve	Brier score
Logistic regression	0.83	0.82-0.88	<.001	0.37	0.06
Random forest	0.85	0.83-0.88	<.001	0.39	0.06
Gradient boosting machine	0.88	0.85-0.90	<.001	0.50	0.05
Feedforward neural network	0.85	0.83-0.88	<.001	0.41	0.06

Calculating the Importance of Feature Variables in Mortality-Prediction Models

The rank of important variables in the models for predicting 30-day mortality is presented in Table 5. In LR, systolic blood pressure was identified as the most important variable. RF indicated serum aspartate aminotransferase as important, while GBM and FNN indicated serum protein and serum phosphorus important, respectively. Overall, vital signs and several laboratory values such as arterial blood pH, O_2, and CO_2 concentration were mainly identified as important variables across the different ML methods.

Table 5. Top 10 important variables of each machine learning model.

Rank	Logistic regression	Random forest	Gradient boosting machine	Feedforward neural network
1	Systolic blood pressure	Serum aspartate aminotransferase	Serum protein	Serum phosphorus
2	Diastolic blood pressure	Pa_{CO2}	Age	Pa_{CO2}
3	Respiratory rate	Arterial pH	Serum phosphorus	Hemoglobin
4	Pa_{CO2}	Pa_{O2}	Systolic blood pressure	Systolic blood pressure
5	Arterial pH	Serum alanine aminotransferase	Platelet	Normal sinus rhythm in electrocardiogram
6	Pa_{O2}	Total bilirubin	Serum aspartate aminotransferase	Estimated glomerular filtration rate
7	Aspartate aminotransferase	Creatine kinase-myocardial band	Pa_{O2}	Serum glucose
8	Pulse rate	White blood cell	Serum albumin	Platelet
9	Blood urea nitrogen	Serum sodium	Pulse rate	Pa_{O2}
10	Serum phosphorus	Platelet	Activated partial thromboplastin time	Arterial pH

Discussion

Principal Findings

This was a retrospective study that applied ML to structured and unstructured patient data from the EMR of a large quaternary hospital to develop a risk prediction model for 30-day adverse cardiac events in patients who underwent PCI or CABG. We comparatively evaluated the performance of several models; all models demonstrated outstanding results with AUROCs more than 0.90 with excellent calibration. On external validation, the performance in predicting 30-day mortality decreased; however, it remained favorable. Dynamic time-series data, including laboratory values, vital signs, and medications, demonstrated the best performances, which mainly contributed to outstanding performance of the models.

Traditional risk prediction models are derived from a small set of selected risk factors based on the significant univariate relationship with the end point on LR, which might deteriorate the predictive performance. Moreover, it is difficult to include new and more discriminatory risk factors into the traditional models, which limits their extension ability [12]. Advances in big data solutions allow for storage, management, and mining of large volumes of structured and semistructured data such as complex health care data [22]. Along the emergence of big data, ML provides an alternative approach to establish prediction modeling that might address the current limitations. In this context, we aimed to develop and validate ML models by using longitudinal and heterogeneous data of various EMR parameters to predict mortality or MACE at 30 days after PCI or CABG. In addition, we explored a general framework for constructing models by categorizing the data set into static basic data, dynamic time-series data, and disease-specific data to examine the potential applicability. This study revealed encouraging results, which indicate that ML-based models for predicting adverse events after invasive coronary treatment might be feasible and applicable as a clinical decision supporting system in hospitals with fully implemented EMR protocols. Furthermore, this approach can be extended to various disease entities or clinical events for improvement in quality of care and patient outcomes.

In this study, we found that the algorithms developed from a large single-center EMR database were reliable for use in the population of a different hospital, albeit with a relatively low performance. Of note, different hospitals serve dissimilar patient populations and have divergent clinical practice patterns; therefore, the EMR data reflecting the real-world clinical practice in each hospital has its own distinct characteristics. Hence, a somewhat low performance of the proposed prediction models in a different cohort can be anticipated. Ideally, a model that achieves the highest possible level of generalizability is desirable. However, there have been concerns about whether a model developed at 1 center can be applied to another center [9]. In medicine, there are too many practice patterns and other local idiosyncrasies that make learning a broadly applicable model effectively difficult [23,24]. In respect that the ultimate application of prediction models built with EMR data is integration with the clinical decision support system for personalized medicine, optimizing individual centers' particular prediction model may be more important rather than extending generalizability. Hence, although the developed algorithms from a single-center EMR database can be used with the database of a different hospital, individual prediction models based on the EMR data of each single hospital would be preferable for highly optimized performance.

Predictive models with EMR data frequently rely on structured data. However, given the volume and richness of data available in unstructured clinical notes or reports, ML models might benefit from leveraging text mining tools to enhance the model [22,25]. Hence, we text-mined various cardiac-specific data such as image and functional studies and detailed information about PCI and CABG, although the process required diverse strategies and tasks. In this study, text-mined cardiac-specific data showed a fair ability to predict 30-day mortality risk. Although valuable, there are still some challenges in applying text-mined data in ML, particularly owing to the vagueness, impreciseness, and uncertain clinical information in EMR data [12]. In contrast, utilizing structured data is simple if the database and automated process system for extraction, transformation, and loading of data are well-established. Algorithms with only time-series dynamic data, which is the typical large-volume structured data, outperformed and primarily contributed to the excellent final results. Intuitively, it is believed that the learning model will perform better if more data are integrated into learning [26]. Our results indicate that using only large amount of reliable structured data of EMR could offer an opportunity to develop proper risk prediction models. However, although improvement in clinical data collection processes is necessary, fundamentally, significant clinical information should be recorded digitally in a cohesive and standardized manner in the EMR system.

Limitations

Several limitations of this study should be noted. First, the cardiovascular event rates, including mortality, might be underestimated because events were captured only from a single-center EMR database. Linking it with the national claim data or health insurance data might possibly capture the events more accurately. Second, although ease of interpretation is vital for evaluation of the models [27,28], the black box nature of ML makes it difficult to be used in health care. Hence, we tried to assess the importance of the variables through several experiments; however, there is still a lack of "explainability" of the prediction models. For ML methods to be readily adopted in real-world clinical practice, they must be interpretable without compromising on accuracy [29]. Future works focusing on developing explainable ML models are necessary to provide tailored feedback to physicians. Third, other ML methods such as recurrent neural networks, which have shown advantage in leveraging the dynamic features, were not investigated in this work; this needs to be explored in future studies [26,29]. Fourth, although EMR data within 1 year before index procedures were used for all populations, different EMR follow-up times prior to index procedures were not taken into account to develop models. Finally, we did not conduct external validation for MACE. Because physician adjudication of myocardial infarction, stroke, or repeat revascularization events is resource-intensive and time-consuming in a large-scaled record cohort, comprehensive source reviews and final ascertainment were substantially challenged. In order to expand the use of the EMR-based ML approach, optimization for computerized detection and adjudication of clinical outcomes will require considerable investment of time and collaboration with institutional information technology and bioinformatics professionals.

Conclusion

Exploiting the diverse parameters of EMR data sets, we developed and validated ML models for predicting the 30-day mortality risk following PCI or CABG. The ML algorithms showed excellent performance, and the applied framework can be generalized for various health care prediction models. This study suggests that ML using the real-word clinical data set can provide a substantial method of developing risk prediction models. Future studies are warranted to establish the clinical effectiveness of this approach and real-time application at the point of care.

Acknowledgments

This work was supported by the Institute for Information and Communications Technology Promotion grant funded by the Korean government (Ministry of Science and Information Communications Technology; 2018-0-00861, Intelligent Software Technology Development for Medical Data Analysis) and the Korea Medical Device Development Fund grant funded by the Korea government (Ministry of Science and Information Communications Technology, Ministry of Trade, Industry and Energy, Ministry of Health and Welfare, Ministry of Food and Drug Safety; Project: KMDF_PR_20200901_0097).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures and tables.
[DOCX File , 145 KB - medinform_v10i5e26801_app1.docx]

Multimedia Appendix 2
Time-series analysis.
[DOCX File , 21 KB - medinform_v10i5e26801_app2.docx]

References

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, American Heart Association Council on EpidemiologyPrevention Statistics CommitteeStroke 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]
2. Morrow DA. Cardiovascular risk prediction in patients with stable and unstable coronary heart disease. Circulation 2010 Jun 22;121(24):2681-2691. [doi: 10.1161/CIRCULATIONAHA.109.852749] [Medline: 20566966]
3. Neumann F, Sousa-Uva M, Ahlsson A, Alfonso F, Banning AP, Benedetto U, et al. 2018 ESC/EACTS Guidelines on myocardial revascularization. EuroIntervention 2019 Feb;14(14):1435-1534. [doi: 10.4244/eijy19m01_01]
4. Weintraub WS, Grau-Sepulveda MV, Weiss JM, Delong ER, Peterson ED, O'Brien SM, et al. Prediction of long-term mortality after percutaneous coronary intervention in older adults: results from the National Cardiovascular Data Registry. Circulation 2012 Mar 27;125(12):1501-1510 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.111.066969] [Medline: 22361329]
5. Farooq V, Brugaletta S, Serruys PW. Contemporary and evolving risk scoring algorithms for percutaneous coronary intervention. Heart 2011 Dec;97(23):1902-1913. [doi: 10.1136/heartjnl-2011-300718] [Medline: 22058284]
6. Karim MN, Reid CM, Cochrane A, Tran L, Alramadan M, Hossain MN, et al. Mortality risk prediction models for coronary artery bypass graft surgery: current scenario and future direction. J Cardiovasc Surg (Torino) 2017 Dec;58(6):931-942. [doi: 10.23736/S0021-9509.17.09965-7] [Medline: 28497663]
7. Wu C, Camacho FT, Wechsler AS, Lahey S, Culliford AT, Jordan D, et al. Risk score for predicting long-term mortality after coronary artery bypass graft surgery. Circulation 2012 May 22;125(20):2423-2430 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.111.055939] [Medline: 22547673]
8. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. Curr Epidemiol Rep 2018 Dec;5(4):331-342 [FREE Full text] [doi: 10.1007/s40471-018-0165-9] [Medline: 30555773]
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
10. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 2017 Jun 14;38(23):1805-1814 [FREE Full text] [doi: 10.1093/eurheartj/ehw302] [Medline: 27436868]
11. Huang Z, Chan T, Dong W. MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records. J Biomed Inform 2017 Feb;66:161-170 [FREE Full text] [doi: 10.1016/j.jbi.2017.01.001] [Medline: 28065840]
12. Hu D, Dong W, Lu X, Duan H, He K, Huang Z. Evidential MACE prediction of acute coronary syndrome using electronic health records. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):61 [FREE Full text] [doi: 10.1186/s12911-019-0754-7] [Medline: 30961585]
13. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One 2018;13(8):e0202344 [FREE Full text] [doi: 10.1371/journal.pone.0202344] [Medline: 30169498]
14. Hernesniemi JA, Mahdiani S, Tynkkynen JA, Lyytikäinen LP, Mishra PP, Lehtimäki T, et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study. Ann Med 2019 Mar;51(2):156-163 [FREE Full text] [doi: 10.1080/07853890.2019.1596302] [Medline: 31030570]
15. Shin S, Kim WS, Lee J. Characteristics desired in clinical data warehouse for biomedical research. Healthc Inform Res 2014 Apr;20(2):109-116 [FREE Full text] [doi: 10.4258/hir.2014.20.2.109] [Medline: 24872909]
16. Ahn I, Na W, Kwon O, Yang DH, Park G, Gwon H, et al. CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases. BMC Med Inform Decis Mak 2021 Jan 28;21(1):29 [FREE Full text] [doi: 10.1186/s12911-021-01392-2] [Medline: 33509180]
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 07;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]
18. Hicks KA, Mahaffey KW, Mehran R, Nissen SE, Wiviott SD, Dunn B, Standardized Data Collection for Cardiovascular Trials Initiative (SCTI). 2017 Cardiovascular and Stroke Endpoint Definitions for Clinical Trials. J Am Coll Cardiol 2018 Mar 06;71(9):1021-1034 [FREE Full text] [doi: 10.1016/j.jacc.2017.12.048] [Medline: 29495982]
19. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Aug;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
20. Amarasingham R, Velasco F, Xie B, Clark C, Ma Y, Zhang S, et al. Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models. BMC Med Inform Decis Mak 2015 May 20;15:39 [FREE Full text] [doi: 10.1186/s12911-015-0162-6] [Medline: 25991003]

21. Lipton Z, Kale D, Wetzel R. Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series. 2016 Presented at: Machine Learning for Healthcare Conference; August 19-20; Children's Hospital LA, Los Angeles, CA, USA.
22. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
23. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020 Sep;2(9):e489-e492 [FREE Full text] [doi: 10.1016/S2589-7500(20)30186-2] [Medline: 32864600]
24. Kakarmath S, Golas S, Felsted J, Kvedar J, Jethwani K, Agboola S. Validating a Machine Learning Algorithm to Predict 30-Day Re-Admissions in Patients With Heart Failure: Protocol for a Prospective Cohort Study. JMIR Res Protoc 2018 Sep 04;7(9):e176 [FREE Full text] [doi: 10.2196/resprot.9466] [Medline: 30181113]
25. Akbilgic O, Homayouni R, Heinrich K, Langham M, Davis R. Unstructured Text in EMR Improves Prediction of Death after Surgery in Children. Informatics 2019 Jan 10;6(1):4. [doi: 10.3390/informatics6010004]
26. 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]
27. Guo W, Ge W, Cui L, Li H, Kong L. An Interpretable Disease Onset Predictive Model Using Crossover Attention Mechanism From Electronic Health Records. IEEE Access 2019;7:134236-134244. [doi: 10.1109/access.2019.2928579]
28. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
29. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR Workshop Conf Proc 2016 Aug;56:301-318 [FREE Full text] [Medline: 28286600]

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CABG: coronary artery bypass grafting
EMR: electronic medical record
FNN: feedforward neural network
GBM: gradient boosting machine
LR: logistic regression
MACE: major adverse cardiac events
ML: machine learning
PCI: percutaneous coronary intervention
RF: random forest
SPECT: single-photon emission computed tomography

Edited by C Lovis; submitted 28.12.20; peer-reviewed by N Spartano, D Hu; comments to author 16.02.21; revised version received 10.06.21; accepted 31.01.22; published 11.05.22.

Please cite as:
Kwon O, Na W, Kang H, Jun TJ, Kweon J, Park GM, Cho Y, Hur C, Chae J, Kang DY, Lee PH, Ahn JM, Park DW, Kang SJ, Lee SW, Lee CW, Park SW, Park SJ, Yang DH, Kim YH
Electronic Medical Record–Based Machine Learning Approach to Predict the Risk of 30-Day Adverse Cardiac Events After Invasive Coronary Treatment: Machine Learning Model Development and Validation
JMIR Med Inform 2022;10(5):e26801
URL: https://medinform.jmir.org/2022/5/e26801
doi:10.2196/26801
PMID:35544292

Original Paper

Exploring Sentiment and Care Management of Hospitalized Patients During the First Wave of the COVID-19 Pandemic Using Electronic Nursing Health Records: Descriptive Study

Juan Nicolás Cuenca-Zaldívar[1,2*], PT, MSc, PhD; Maria Torrente-Regidor[3,4*], MD, MSc, PhD; Laura Martín-Losada[1,2*], RN, MSc; César Fernández-De-Las-Peñas[5*], PT, MSc, PhD; Lidiane Lima Florencio[5*], PT, MSc, PhD; Pedro Alexandre Sousa[6*], BE, MSc, PhD; Domingo Palacios-Ceña[7*], RN, MSc, PhD

[1]Research Group in Nursing and Health Care, Puerta de Hierro Health Research Institute - Segovia de Arana, Majadahonda, Spain

[2]Functional Recovery Unit, Guadarrama Hospital, Guadarrama, Spain

[3]Servicio de Oncología Médica, Hospital Universitario Puerta de Hierro, Majadahonda, Spain

[4]Faculty of Health Sciences, Universidad Francisco de Vitoria, Majadahonda, Spain

[5]Research Group of Manual Therapy, Department of Physical Therapy, Occupational Therapy, Physical Medicine, and Rehabilitation, Universidad Rey Juan Carlos, Alcorcón, Spain

[6]Department of Electrical Engineering, Faculty of Science and Technology, Universidade Nova de Lisboa, Lisbon, Portugal

[7]Research Group of Humanities and Qualitative Research in Health Science, Department of Physical Therapy, Occupational Therapy, Physical Medicine and Rehabilitation, Universidad Rey Juan Carlos, Alcorcón, Spain

[*]all authors contributed equally

Corresponding Author:
Juan Nicolás Cuenca-Zaldívar, PT, MSc, PhD
Research Group in Nursing and Health Care
Puerta de Hierro Health Research Institute - Segovia de Arana
C Joaquín Rodrigo, 1
Majadahonda, 28222
Spain
Phone: 34 639962935
Email: jcuenzal@yahoo.es

Abstract

Background: The COVID-19 pandemic has changed the usual working of many hospitalization units (or wards). Few studies have used electronic nursing clinical notes (ENCN) and their unstructured text to identify alterations in patients' feelings and therapeutic procedures of interest.

Objective: This study aimed to analyze positive or negative sentiments through inspection of the free text of the ENCN, compare sentiments of ENCN with or without hospitalized patients with COVID-19, carry out temporal analysis of the sentiments of the patients during the start of the first wave of the COVID-19 pandemic, and identify the topics in ENCN.

Methods: This is a descriptive study with analysis of the text content of ENCN. All ENCNs between January and June 2020 at Guadarrama Hospital (Madrid, Spain) extracted from the CGM Selene Electronic Health Records System were included. Two groups of ENCNs were analyzed: one from hospitalized patients in post–intensive care units for COVID-19 and a second group from hospitalized patients without COVID-19. A sentiment analysis was performed on the lemmatized text, using the National Research Council of Canada, Affin, and Bing dictionaries. A polarity analysis of the sentences was performed using the Bing dictionary, SO Dictionaries V1.11, and Spa dictionary as amplifiers and decrementators. Machine learning techniques were applied to evaluate the presence of significant differences in the ENCN in groups of patients with and those without COVID-19. Finally, a structural analysis of thematic models was performed to study the abstract topics that occur in the ENCN, using Latent Dirichlet Allocation topic modeling.

Results: A total of 37,564 electronic health records were analyzed. Sentiment analysis in ENCN showed that patients with subacute COVID-19 have a higher proportion of positive sentiments than those without COVID-19. Also, there are significant differences in polarity between both groups (Z=5.532, P<.001) with a polarity of 0.108 (SD 0.299) in patients with COVID-19 versus that of 0.09 (SD 0.301) in those without COVID-19. Machine learning modeling reported that despite all models presenting

high values, it is the neural network that presents the best indicators (>0.8) and with significant *P* values between both groups. Through Structural Topic Modeling analysis, the final model containing 10 topics was selected. High correlations were noted among topics 2, 5, and 8 (pressure ulcer and pharmacotherapy treatment), topics 1, 4, 7, and 9 (incidences related to fever and well-being state, and baseline oxygen saturation) and topics 3 and 10 (blood glucose level and pain).

Conclusions: The ENCN may help in the development and implementation of more effective programs, which allows patients with COVID-19 to adopt to their prepandemic lifestyle faster. Topic modeling could help identify specific clinical problems in patients and better target the care they receive.

(JMIR Med Inform 2022;10(5):e38308) doi:10.2196/38308

KEYWORDS

electronic health records; COVID-19; pandemic; content text analysis

Introduction

On March 11, 2020, the World Health Organization declared COVID-19 a global pandemic [1]. SARS-CoV-2 presented a great capacity for contagion, spread, and high mortality, which collapsed health care systems worldwide [2,3]. Owing to the sudden spread of the virus, health care professionals have undergone a huge, rapid, and profound change in their professional workplace to combat COVID-19.

In this context, receiving a diagnosis of COVID-19 and being admitted to hospital, often in intensive care units for periods of weeks or even months, provoked a sense of helplessness and near death. This situation has led to an increased prevalence of mental health problems owing to a high rate of prevalence of anxiety and depression among patients with COVID-19 [4], which approached 30% [5] and led to posttraumatic stress in up to 96.2% of those affected [6]. Medical activity has focused primarily on the treatment of the disease [7] and research has focused on epidemiological [8,9], clinical, and pathophysiological factors [10,11].

During the COVID-19 pandemic, electronic health records (EHRs) have provided an agile response to the needs of health care workers and researchers through useful data exploitation [12,13] by presenting information quickly and efficiently, for primary and secondary uses in clinical care [14]. This system allowed having complete and coherent information regardless of where or by whom it was generated, enabling it to follow the timeline of the patient's disease, including symptoms, acute events, or changes in their treatment or health status [15], which was especially key given the high rotation of health care workers. On the other hand, a fundamental point in EHRs is correct recording of the information in order to be able to make effective and safe clinical decisions for the patient [16]. Previous studies show how the lack of registration of information on the diagnostic process, identification, and listing of events on care and treatment can affect the monitoring of the quality, safety, and efficacy of health care interventions [17]. These clinical notes can be only written by EHR users responsible for patient care, such as doctors, nurses, and assistant nurses [18,19]. Wisner et al [20] showed that the absence or limitation in nursing clinical narratives, comments, and clinical notes hinders clinical reasoning and decision-making along with the transmission of information between the different shifts.

Electronic nursing clinical notes (ENCN) are documents in which nurses describe health status, nursing care, medication, and other observations about patients [21]. In these texts, they also describe their observations and opinions in an attempt to better understand the patients' condition and opinions [22] and among them, the feelings perceived during their interaction [21]. The appropriate use of ENCN can help improve both physical and mental health care of hospitalized patients [23].

Much of the relevant information is recorded in ENCN in the form of free text (unstructured), known as clinical notes, which makes analysis and decision-making very difficult. This has stimulated the development of semantic analysis methods [24,25] that allow in-depth exploration of the clinical information potentially available in health services [26] and determine the amount of information collected about a clinical process or condition [18-20] and the content of that information regarding specific topics.

Sentiment or opinion analysis allows the analysis of positive or negative sentiments in a text by using precalibrated dictionaries of terms [27]. Polarity facilitates the qualification of these sentiments in the context of sentences; for example, the term "happy" denotes a clearly positive sentiment, but if it is preceded by "not happy" in the sentence, the polarity is reversed toward a negative value [28]. Sentiment analysis in the health care domain has been used in the analysis of social networks [29,30], suicide notes [31], or radiology notes [32], as well as nursing notes [33]. The application of this type of analysis provides insight into patients' attitudes toward the contextual polarity of ENCN and assesses symptoms related to their mental health, which may not have been detected through direct analysis [22].

Latent Dirichlet Allocation (LDA) thematic pattern analysis is a technique to detect hidden topics in a corpus of texts [34]. It assumes topics with word clusters in which the distribution of words within each topic is taken into account, along with the distribution of topics throughout the corpus [35]. This technique has been used in social network analysis, news [36,37], or in response to government policies [38]. Biomedical terms have been found to form specific topics [39,40]; so, this analysis can provide useful clinical information [35].

To our best of knowledge, there are currently no studies describing the use of ENCN for the determination of sentiment and polarity (rejection-acceptance) as well as the identification of clinical practices of interest of hospitalized patients during the start of the first wave of the COVID-19 pandemic.

Therefore, the objectives of this study were the following: (1) analysis of patient´s sentiments through the analysis of the free text of the ENCN, (2) comparison of the sentiments and polarity of hospitalized patients in post–intensive care units for COVID-19 with those hospitalized in non–COVID-19 wards, (3) temporal analysis of the patients´ sentiments during the first wave of the pandemic (January to June 2020) through the ENCN, and (4) identification of the contents and topics that appear in the ENCN.

Methods

Design

This is a descriptive study that involves an analysis of the textual content of the ENCN [41]. Through the analysis of narrative texts, the positive and negative sentiments of patients can be described and analyzed [42]. The object of the textual analysis studies is to understand how a certain event affects the attitudes and behaviors of people. This study focuses on the ENCN of the nurses who worked during an outbreak of the COVID-19 pandemic in a Spanish hospital [41].

Ethical Considerations

This study was approved by the Clinical Research Ethics Committee at the Hospital Universitario Puerta de Hierro Majadahonda de Madrid (07/400080.9/22). Also, for reviewing clinical histories and data, we had approval from the Guadarrama Hospital Center Management. At all times, the confidentiality of the information was preserved, thus ensuring responsible use of the data, as established by current Spanish regulations and in accordance with the tenets of the Declaration of Helsinki.

Setting, Sample, and Data Collection Tools

All clinical notes contained in the ENCN registered between January and June 2020 at Guadarrama Hospital were extracted from the CGM Selene EHR System (CompuGroup Medical Deutschland AG). Guadarrama Hospital is a mid-term stay hospital in the Community of Madrid, with 144 beds, and provides rehabilitation and long-term care to patients with chronic pathologies; however, during the COVID-19 pandemic, it also provided care to patients with a COVID-19 infection.

The analyzed records collect follow-up data from the day of admission until discharge or death, collecting up to 3 records per day in each work shift (morning shift from 8 AM to 3 PM, afternoon shift from 3 PM to 10 PM, and night shift from 10 PM to 8 AM). ENCN from two groups of nurses were analyzed: one from nurses working with hospitalized post–intensive care unit patients with COVID-19 and a the other from nurses working in non–COVID-19 wards. The hospital´s physicians diagnosed and confirmed COVID-19 and assigned patients to the different wards.

Statistical Analysis

For the statistical analysis, the R package (version 3.5.1; R Foundation for Statistical Computing) was used. The level of significance was established at $P<.05$.

Sentiment Analysis

Previously, the text was standardized by lemmatizing it and cleaning up the stop words. A sentiment analysis was performed on the text using the National Research Council of Canada's (NRC's) Emotion Lexicon [43], Affin [44], and Bing [45] dictionaries. All three of these lexicons are based on unigrams or single Spanish words that assign scores for positive or negative sentiment. In addition, the NRC dictionary categorizes words into emotional categories of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, while the Affin lexicon assigns words with a score between –5 and +5, with negative values indicating negative sentiment and positive values indicating positive sentiment. The presence of significant differences between ENCN in groups of patients with and those without COVID-19 was verified using the Pearson chi-square test, with Bonferroni correction for post hoc analysis. The temporal evolution of sentiments in both groups was evaluated using the Dynamic Time Warp test, which allows comparing time series of different lengths using the normalized Euclidean distance.

Polarity Analysis

In addition, a polarity analysis (Textboxes 1 and 2) of the sentences was performed using the Bing dictionary, the SO Dictionaries V1.11, and Spa [46-48] dictionary as amplifiers and decrementators, and those proposed by Vilares et al [49] as deniers. The Mann-Whitney *U* test was used between the two groups of patients to test significant differences after verifying the nonnormal distribution of polarity using the Kolmogorov-Smirnov test with Lilliefors correction.

Textbox 1. The polarity calculation process.

Four phases were used progressively for the analysis of acceptance-rejection (polarity):

Phase 1. We created a file with the text of the interviews broken down by phrases for textual analysis.

Phase 2. We calculated polarity using the Bing Sentiment Dictionary, the amplifiers and deamplifiers from SO Dictionaries V1.11 and Spa, and the negators proposed by Vilares et al [49].

Phase 3. We calculated the scatterplot of the sentences in the text regarding neutrality to identify positive or negative trends.

Phase 4. The evolution of the emotional valence (positive-negative) would be shown throughout the interviews. We applied Fourier transformation to confirm the polarity trend.

Textbox 2. Formula and dictionaries used to calculate polarity.

The analysis was carried out using the Bing dictionary [28]. The Bing dictionary determines the positivity (acceptance) or negativity (rejection) of each word used. Also, the amplifiers and deamplifiers of SO Dictionaries V1.11 and Spa dictionary [29-31] were used, along with negators proposed by Denecke et al [32].

To calculate the polarity (δ), a context cluster of words (x^T_i) is formed around each polarized word using the Bing dictionary [28], taking by default 4 words before and 2 words after it (if there is any comma in the cluster, it will only include the words that are after the comma), and those will be treated as valence shifters.

The words in this cluster are labeled as neutral (x^0_i), negators (x^N_i), amplifiers (x^a_i) or de-amplifiers (x^d_i) using the dictionary SO Dictionaries V1.11 Spa2 [47] and the negators proposed by Hu and Liu [48]. Neutral words do not add to the equation but affect the word count (n).

Each polarized word (negative or positive) is weighted (w) on the basis of the context cluster weights (x^T_i) and further weighted by the number and position of the valence shifters directly surrounding it. A weight (c) can be added and applied to both amplifiers and deamplifiers (with a default value of 0.8 and a lower limit for the deamplifiers of –1).

Finally, the context cluster (x^T_i) is added and divided by the square root of the number of words (√n) to generate a polarity score (δ) that, by default, is not limited in value.

The final result is the following formula:

Where:

ENCN Comparison

Machine learning enables the automation of large amounts of text by model training [50]. Machine learning techniques were applied in order to evaluate the presence of significant differences in the ENCN among patients with and those without COVID-19. For this, the models were created on a random subsample of 75% of the text, applying them to the remaining 25%. The applied models were Support Vector Machine, Naive-Bayes, random forest, and neural network. The quality of the models was evaluated using the area under the curve (AUC), sensitivity and specificity, the κ index, and accuracy with its level of significance. Values above 0.8 and significant *P* values ($P<.05$) were considered the cutoff point.

Topic and Content Analysis

A structural analysis of thematic models (STM) was performed to study the abstract topics that occur in the comments, using LDA topic modeling but allows their inclusion as covariates in the model, the temporal evolution, and the presence of the of ENCN in groups of patients with subacute COVID-19 and those without COVID-19 [33]. The optimal number of topics was determined while considering exclusivity [34] and semantic coherence [35] as criteria. Exclusivity evaluates if the top words for the topics appear within top words of other topics, while semantic coherence shows if the words that are most associated with the corresponding themes occur equally within the documents; in both cases, higher values are better. The effect of the topics of the final model between ENCN in patients with and those without COVID-19 was analyzed, along with the temporal evolution in the prevalence of the appearance of global themes between both groups. The interaction graph was used to determine the presence of significant differences in the evolution of prevalence between both groups. An analysis of the content of the topics and the differences in themes between both groups was carried out, while the network graph allowed for the detection of the presence of categories between topics.

Results

A total of 37,564 records were analyzed, after eliminating 24,101 duplicates (ie, ENCN that had been copied and pasted from previous ones). ENCN were produced by 77 nurses distributed by working shift, hospital unit, and months (Table 1).

These records correspond to 710 patients, whose baseline demographics and clinical data are shown depending on whether or not they were infected with SARS-CoV-2 (sociodemographic data in Multimedia Appendix 1).

Table 1. Distribution of electronic clinical nursing notes by working shift, units, and time (in months).

	Electronic nursing clinical notes for the COVID-19 group, n (%)	Electronic nursing clinical notes for the non–COVID-19 group, n (%)
Working shift		
Morning	5161 (13.7)	10,791 (28.7)
Afternoon	3637 (9.6)	7931 (21.1)
Night	3992 (10.6)	6050 (16.1)
Month		
January	161 (0.4)	7360 (19.5)
February	466 (1.2)	6225 (16.5)
March	2457 (6.5)	4104 (10.9)
April	5467 (14.5)	396 (1.0)
May	3245 (8.6)	2093 (5.5)
June	994 (2.6)	4594 (12.2)

Sentiment Analysis

The differences in the sentiments expressed in the ENCN between both groups were significant in the NRC dictionary (χ^2_9=360.6, *P*<.001), Afinn lexicon both in the scores (χ^2_8=385.3, *P*<.001) and polarity (χ^2_1=232.7, *P*<.001), and Bing dictionary (χ^2_1=368.9, *P*<.001). Post hoc tests showed significant differences among all levels (Multimedia Appendices 2 and 3).

In the ENCN of patients with COVID-19, there is a higher proportion of positive sentiments than that in the non–COVID-19 group. The most frequently expressed emotion is sadness, which was greater in the non–COVID-19 group, followed by trust, which appears to be similar in both groups. Sentiments with negative scores (–2) are more frequent in the non–COVID-19 group, while that of positive sentiments was higher in the COVID-19 group (+2) (Table 2).

The evolution of the sentiments expressed in the ENCN was similar in both groups, revealing a drastic reduction during April and May in the non–COVID-19 group, consistent with the peak of the pandemic (Multimedia Appendix 4).

However, higher values were generally observed in the sentiments expressed in the COVID-19 group when they were analyzed with the Afinn dictionary, where the emotional valences doubled those of patients without COVID-19 and where we observed a clear asymmetry in the distribution of the most negative sentiments (scores of –5).

The distances between both time series are generally small; that is, <0.2. The NRC dictionary showed the greatest differences between the 2 groups in the emotions of surprise and sadness, in the positive sentiments of the Bing dictionary, and in the negative ones of the Afinn dictionary (Multimedia Appendix 5).

Table 2. Sentiment scores by dictionary.

Dictionaries	Electronic nursing clinical notes for the COVID-19 group, mean	Electronic nursing clinical notes for the non–COVID-19 group, mean	*P* value
National Research Council of Canada dictionary			<.001
Anger	2.8	2.7	
Anticipation	8.7	8.5	
Disgust	3.3	3.9	
Fear	6.6	7.3	
Joy	5.2	4.6	
Sadness	17.1	18.0	
Surprise	3.6	2.9	
Trust	10.6	10.4	
Negative	20.0	21.3	
Positive	22.0	20.6	
Afinn dictionary			<.001
–5	0.0	0.0	
–4	0.1	0.1	
–3	2.1	1.7	
–2	36.0	43.4	
–1	18.7	18.3	
+1	6.2	6.6	
+2	36.0	29.2	
+3	0.9	0.7	
+4	0.0	0.0	
Afinn dictionary (positive-negative)			<.001
Negative	56.9	63.5	
Positive	43.1	36.5	
Bing dictionary			<.001
Negative	34.3	40.4	
Positive	65.7	59.6	

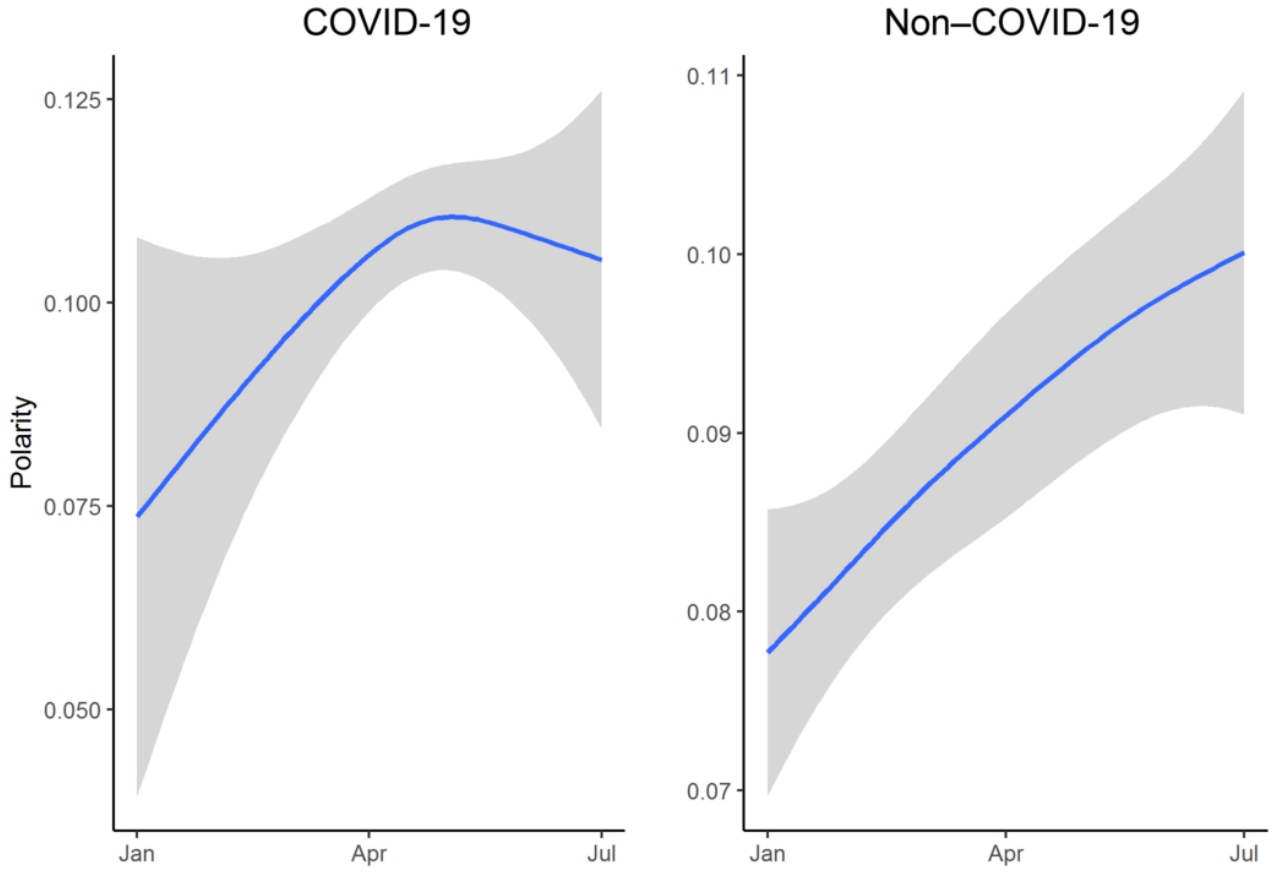
Polarity Analysis

Polarity scores were nonnormally distributed between the COVID-19 and non–COVID-19 groups (P<.001).

There are significant differences in polarity between both groups (Z=5.532, P<.001): 0.108 (SD 0.299) in patients with COVID-19 versus 0.09 (SD 0.301) in those without COVID-19.

When both groups were compared, we verified how the polarity presents a clear upward trend in ENCN of the non–COVID-19 group, while in ENCN of the COVID-19 group, the most positive value was attained in April to decrease later with higher values than those of the non–COVID-19 group (Figure 1).

Figure 1. Polarity of patients' comments.



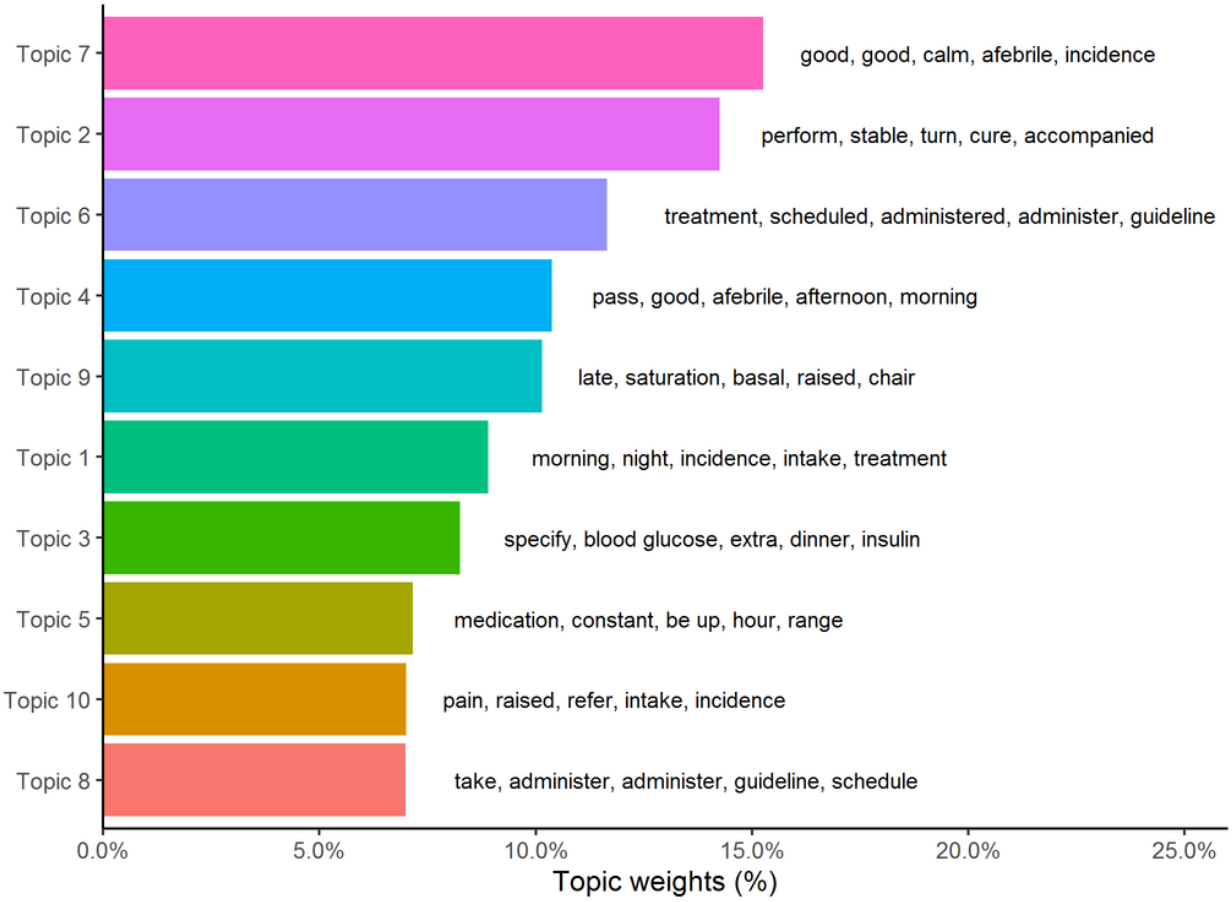
STM

The selected model contains 10 topics. The topics tend to be assigned to a few comments, which indicates a high specificity in their content. The presence of the following concepts was hypothesized on the basis of the selected topic weights (see Textbox 3 and Figure 2)

Textbox 3. Topics identified from electronic nursing clinical notes.

Topic 1: Incidents in each working shift.
Topic 2: Application of pressure ulcer treatments.
Topic 3: Blood glucose level and insulin pattern.
Topic 4: Presence or absence of fever in relation to general condition.
Topic 5: Pharmacotherapy treatment and vital signs control.
Topic 6: Administration of the treatment schedule.
Topic 8: Taking the medication.
Topic 7: Incidents that affect the general well-being of the patient.
Topic 9: Baseline oxygen saturation.
Topic 10: Incidents related to the appearance of pain.

Figure 2. Topic weights and the 5 most frequent words by topic.



An increase in the prevalence of topics was observed in the second half of the semester, which coincided with the time of admission of patients with COVID-19. Over time, patients with COVID-19 showed a higher prevalence of items 9 (baseline oxygen saturation) and 7 (general well-being), with a lower proportion of items 2 (pressure ulcer treatments), 3 (insulin), 5 (drug therapy and vital sign control), and 10 (pain control) than those without COVID-19. Our findings reported that the main problems among patients with COVID-19 were related to initial oxygen saturation and general well-being, while in those without COVID-19, problems were related to pressure ulcer treatment, pain, diabetes, and drug therapy. The analysis shows different nuances between patients with and those without COVID-19 in the topics of the model. Control of baseline oxygen saturation, blood glucose level, and ingestion, as well as fever, are of greater importance to patients with COVID-19; while among those without COVID-19, pain, insulin dose, pressure ulcer treatment, and pharmacotherapy were the priority topics. In both groups, there is a common concern for the general condition and well-being of the patients, as well as for the control of the treatment regimen.

These differences are significant between both groups over time, as shown in the interaction graph, with an increase in the proportion of topics in the second half of the semester in the COVID-19 group, while in the first half of the semester, this proportion is higher in the non–COVID-19 group. Topics 9 (baseline oxygen saturation) and 2 (pressure ulcer treatment) present the greatest and significant effects between both groups, while topics 8 and 1 do not show any significant effect.

There was a high correlation among topics 2, 5, and 8 (pressure ulcer care, vital sign control, and pharmacotherapy treatment), topics 1, 4, 7, and 9 (incidences related to working shift, fever and well-being state, and baseline oxygen saturation), and topics 3 and 10 (blood glucose level and pain), while topic 6 (administration of treatment schedule) remains uncorrelated.

Machine Learning Modeling

Although all the models show high values, the neural network showed the best indicators (>0.8) and with significant *P* values. The worst model was the random forest model, which was clearly overfitting (Table 3).

This result coincides with the findings of the thematic model analysis and may indicate significant differences in the type of nursing comment based on the presence or absence of a COVID-19 infection, with the neural network showing excellent values of sensitivity and specificity, as well as precision.

Table 3. Machine learning models quality.

Model	Area under the curve	Sensitivity	Specificity	Accuracy (95% CI)	Accuracy (*P* value)
Support vector machine	0.70	0.91	0.50	0.77 (0.76-0.78)	<.001
Random forest	1.0	1.0	1.0	1 (1-1)	<.001
Naive-Bayes	0.80	0.78	0.82	0.79 (0.79-0.80)	<.001
Neural network	0.96	0.99	0.92	0.97 (0.96-0.97)	<.001

Discussion

Principal Findings

Our findings report a higher proportion of positive sentiments among patients with subacute COVID-19 than that of those without COVID-19. Groups also differed on the polarity of their narratives (*P*<.001). Among the machine learning models, the neural network presented the best indicators. In addition, the final STM containing 10 topics with high correlations among topics 2, 5, and 8 (pressure ulcer and pharmacotherapy treatment), topics 1, 4, 7, and 9 (incidences related to fever and well-being state, and baseline oxygen saturation), and topics 3 and 10 (blood glucose level and pain).

Previous studies show the presence of positive sentiments during the pandemic, reflected in gratitude toward health care workers and community support for vulnerable people [51]. Our results show a higher proportion of positive sentiments in the ENCN of the COVID-19 group than that of the non–COVID-19 group. These results are consistent with those reported by Sahoo et al [52] as the patients had been in the intensive care unit for more than 40 days. The authors suggest that patients tend to become progressively more relaxed and that the experience of the ward environment changes, with situations perceived as positive becoming more frequent. The emotion most frequently expressed in ENCN was sadness, which was observed in the non–COVID-19 group. Most patients without COVID-19 were in the functional recovery unit—these patients are characterized as being older adults with a prolonged hospital stay and with comorbidities often associated with physical pain. The feeling of sadness could be related to physical pain, according to Shirai and Soshi [53]. Age is also considered a predisposing factor according to Wu et al [54], where hospitalized older adults are at a higher disposition to sadness.

Among the 10 main topics of the model selected for ENCN, the topics with the greatest weight were the application of treatments for pressure cutaneous lesions in the non–COVID-19 group and baseline oxygen saturation in the COVID-19 group. In both groups of patients, there was a common concern for the general condition and well-being of the patients, as well as for control on the treatment regimen. The relevant issues detected in the ENCN in the COVID-19 group were the stability of vital signs (fever and oxygen saturation), glucose control, and diet. The importance of oxygen saturation is justified by the respiratory involvement by SARS-CoV-2 infection [55]. Glucose control could be explained by its relationship with diabetes mellitus being a metabolic syndrome considered as high risk with respect to COVID-19 severity; it may also be related to the use of corticosteroids for the anti-inflammatory treatment of respiratory infection [56]. Regarding diet, the frequency of ENCN could be associated with irregular or low intakes due to the acute phase, with anosmia and ageusia being typical symptoms of SARS-CoV-2 infection [55].

In the ENCN in the non–COVID-19 group, the presence of skin lesions as a topic of interest could be explained by the prevalence of dependence in hospitalized patients, the rate of which is 8.7% in Spain. Furthermore, pressure injuries account for 7%, according to the fifth Spanish National Study of Prevalence of pressure ulcers and other chronic wounds [57]. In addition, patients in the non–COVID-19 group present risk factors for skin lesions, such as advanced age, comorbidity, prolonged hospitalization, functional limitations, and urinary incontinence [58]. Other topics of interest in the ENCN for the non–COVID-19 group were insulin dose and pharmacotherapy. The presence of comorbidities, such as diabetes mellitus, is a common concern for nurses in both groups. In Spain, this disease has a prevalence of 12.5% in adults, mostly affecting older adults [59]. Other records referred to the assessment and control of pain, a symptom that is usually associated with rehabilitation processes [56,60].

Text analysis of unstructured ENCNs has been used with success previously to determine the quality of the registry [61] and in other unstructured texts such as patient experience [62]. This type of analysis is considered useful to capture the perception of an event, demonstrating reliability in health sciences and COVID-19 issues [51,62]. The ability to identify new topics of interest and detect areas for improvement is also considered important [63]. Regarding the dictionaries used in this study, all of them (NRC, Affin, and Bing) yielded significant results; hence, the selected words can be considered sensitive and useful in the care of patients with and those without COVID-19.

The application of text mining techniques on clinical text may be a valid source for evaluating the sentiments of hospitalized patients and detecting problems related to their mental health (anxiety, depression, and posttraumatic stress), which may influence the evolution of their illness. These results may help establish early and more effective recovery programs that address these issues and allow those affected to return more quickly to their prepandemic lifestyle.

Finally, topic modeling has made it possible to obtain relevant clinical information from the clinical notes, allowing the identification of clinical problems in providing care to patients with and those without COVID-19, which are clearly differentiated, and which may help guide their care more effectively.

Limitations

This study has limitations. The main outcome could not be compared more broadly owing to the absence of studies on polarity and sentiment in ENCN during the start of first wave of the COVID-19 pandemic. The patients´ sentiments before and during the pandemic could be different; hence, the results of the comparisons between patients with and those without COVID-19 must be interpreted with caution.

Conclusions

ENCN can provide very useful real-time information, identifying the patient's sentiments and their polarity (rejection-acceptance). Additionally, it may serve to identify relevant issues based on the care of different groups of patients, both with and those without COVID-19. This can present an opportunity to direct health care strategies in accordance with the needs detected in hospitalized patients, based on real word data, and may help develop and implement preventive programs.

Acknowledgments

We extend a special thanks to all health care professionals for their work and resilience.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sociodemographic data.
[DOCX File , 14 KB - medinform_v10i5e38308_app1.docx]

Multimedia Appendix 2
Significant differences between sentiments. Post hoc pairwise comparison with the NRC dictionary.
[DOCX File , 13 KB - medinform_v10i5e38308_app2.docx]

Multimedia Appendix 3
Significant differences between all emotional levels. Post hoc pairwise comparison with the Afinn dictionary.
[DOCX File , 13 KB - medinform_v10i5e38308_app3.docx]

Multimedia Appendix 4
Emotions evolution during first semester of 2020.
[DOCX File , 133 KB - medinform_v10i5e38308_app4.docx]

Multimedia Appendix 5
Distances between the time series of ENCN of Covid and non-Covid patients.
[DOCX File , 65 KB - medinform_v10i5e38308_app5.docx]

References

1. Our work. World Health Organization. URL: https://www.who.int [accessed 2020-08-06]
2. Desai AN, Patel P. Stopping the Spread of COVID-19. JAMA 2020 Apr 21;323(15):1516. [doi: 10.1001/jama.2020.4269] [Medline: 32196079]
3. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis 2020 Jul;26(7):1470-1477 [FREE Full text] [doi: 10.3201/eid2607.200282] [Medline: 32255761]
4. Kong X, Zheng K, Tang M, Kong F, Zhou J, Diao L, et al. Prevalence and Factors Associated with Depression and Anxiety of Hospitalized Patients with COVID-19. medRxiv Preprint posted online April 5, 2020. [doi: 10.1101/2020.03.24.20043075]
5. Zhang J, Lu H, Zeng H, Zhang S, Du Q, Jiang T, et al. The differential psychological distress of populations affected by the COVID-19 pandemic. Brain Behav Immun 2020 Jul;87:49-50 [FREE Full text] [doi: 10.1016/j.bbi.2020.04.031] [Medline: 32304883]
6. Bo H, Li W, Yang Y, Wang Y, Zhang Q, Cheung T, et al. Posttraumatic stress symptoms and attitude toward crisis mental health services among clinically stable patients with COVID-19 in China. Psychol Med 2021 Apr;51(6):1052-1053 [FREE Full text] [doi: 10.1017/S0033291720000999] [Medline: 32216863]
7. Jamili S, Ebrahimipour H, Adel A, Badiee Aval S, Hoseini SJ, Vejdani M, et al. Experience of patients hospitalized with COVID-19: A qualitative study of a pandemic disease in Iran. Health Expect 2022 Apr;25(2):513-521 [FREE Full text] [doi: 10.1111/hex.13280] [Medline: 34224643]

8. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun 2020 May;109:102433 [FREE Full text] [doi: 10.1016/j.jaut.2020.102433] [Medline: 32113704]
9. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 2020 Feb 15;395(10223):507-513 [FREE Full text] [doi: 10.1016/S0140-6736(20)30211-7] [Medline: 32007143]
10. Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. J Med Virol 2020 Apr;92(4):441-447 [FREE Full text] [doi: 10.1002/jmv.25689] [Medline: 31994742]
11. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med 2020 Mar 26;382(13):1199-1207 [FREE Full text] [doi: 10.1056/NEJMoa2001316] [Medline: 31995857]
12. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. Brief Bioinform 2021 Mar 22;22(2):812-822 [FREE Full text] [doi: 10.1093/bib/bbaa418] [Medline: 33454728]
13. Reeves JJ, Hollandsworth HM, Torriani FJ, Taplitz R, Abeles S, Tai-Seale M, et al. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. J Am Med Inform Assoc 2020 Jun 01;27(6):853-859 [FREE Full text] [doi: 10.1093/jamia/ocaa037] [Medline: 32208481]
14. Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, Terriza-Torres AI, López-Jiménez EA, Calvo-Boyero F, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on Detailed Clinical Models. J Biomed Inform 2021 Mar;115:103697 [FREE Full text] [doi: 10.1016/j.jbi.2021.103697] [Medline: 33548541]
15. Kim MO, Coiera E, Magrabi F. Problems with health information technology and their effects on care delivery and patient outcomes: a systematic review. J Am Med Inform Assoc 2017 Mar 01;24(2):246-250 [FREE Full text] [doi: 10.1093/jamia/ocw154] [Medline: 28011595]
16. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: An audit of problem list completeness during the COVID-19 pandemic. Int J Med Inform 2021 Jun;150:104452 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104452] [Medline: 33864979]
17. Wright A, McCoy AB, Hickman TT, Hilaire DS, Borbolla D, Bowes WA, et al. Problem list completeness in electronic health records: A multi-site study and assessment of success factors. Int J Med Inform 2015 Oct;84(10):784-790 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.06.011] [Medline: 26228650]
18. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
19. Wallace D, Kecahdi T. Outlier Detection in Health Record Free-Text using Deep Learning. Annu Int Conf IEEE Eng Med Biol Soc 2019 Jul;2019:550-555. [doi: 10.1109/EMBC.2019.8857491] [Medline: 31945959]
20. Wisner K, Lyndon A, Chesla CA. The electronic health record's impact on nurses' cognitive work: An integrative review. Int J Nurs Stud 2019 Jun;94:74-84. [doi: 10.1016/j.ijnurstu.2019.03.003] [Medline: 30939418]
21. Sanglerdsinlapachai N, Plangprasopchok A, Ho TB, Nantajeewarawat E. Improving sentiment analysis on clinical narratives by exploiting UMLS semantic types. Artif Intell Med 2021 Mar;113:102033. [doi: 10.1016/j.artmed.2021.102033] [Medline: 33685589]
22. Chintalapudi N, Battineni G, Canio MD, Sagaro G, Amenta F. Text mining with sentiment analysis on seafarers' medical documents. International Journal of Information Management Data Insights 2021 Apr;1(1):100005 [FREE Full text] [doi: 10.1016/j.jjimei.2020.100005]
23. Lal M, Avatade M, Mudholkar R. Sentiment Analysis and Machine Learning on Clinical Text: An Overview. Paripex Indian J Res 2019;8(6):139-140 [FREE Full text]
24. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform 2015 Aug 13;10(1):183-193 [FREE Full text] [doi: 10.15265/IY-2015-009] [Medline: 26293867]
25. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. Yearb Med Inform 2015 Aug 13;10(1):194-198 [FREE Full text] [doi: 10.15265/IY-2015-035] [Medline: 26293868]
26. Rabaey J, Chandrakasan A, Nikolic B. Digital Integrated Circuits: A Design Perspective (2nd edition). Upper Saddle River, NJ: Prentice-Hall, Inc; 2002.
27. Weissman GE, Ungar LH, Harhay MO, Courtright KR, Halpern SD. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. J Biomed Inform 2019 Jan;89:114-121 [FREE Full text] [doi: 10.1016/j.jbi.2018.12.001] [Medline: 30557683]
28. Stone PJ, Dunphy DC, Smith MS, Ogilvie DM. The General Inquirer: A Computer Approach to Content Analysis. Am J Sociol 1968 Mar;73(5):634-635 [FREE Full text] [doi: 10.1086/224539]
29. Korkontzelos I, Nikfarjam A, Shardlow M, Sarker A, Ananiadou S, Gonzalez GH. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. J Biomed Inform 2016 Aug;62:148-158 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.007] [Medline: 27363901]

30. Ji X, Chun SA, Wei Z, Geller J. Twitter sentiment classification for measuring public health concerns. Soc Netw Anal Min 2015;5(1):13 [FREE Full text] [doi: 10.1007/s13278-015-0253-5] [Medline: 32226558]
31. Pestian JP, Matykiewicz P, Linn-Gust M, South B, Uzuner O, Wiebe J, et al. Sentiment Analysis of Suicide Notes: A Shared Task. Biomed Inform Insights 2012 Jan 30;5(Suppl 1):3-16 [FREE Full text] [doi: 10.4137/bii.s9042] [Medline: 22419877]
32. Denecke K, Deng Y. Sentiment analysis in medical settings: New opportunities and challenges. Artif Intell Med 2015 May;64(1):17-27. [doi: 10.1016/j.artmed.2015.03.006] [Medline: 25982909]
33. Waudby-Smith IER, Tran N, Dubin JA, Lee J. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. PLoS One 2018;13(6):e0198687 [FREE Full text] [doi: 10.1371/journal.pone.0198687] [Medline: 29879201]
34. Sato I. Latent Dirichlet Allocation. Intell Inf 2012;24(4):160. [doi: 10.3156/jsoft.24.4_160_1]
35. Gupta A, Aeron S, Agrawal A, Gupta H. Trends in COVID-19 Publications: Streamlining Research Using NLP and LDA. Front Digit Health 2021;3:686720 [FREE Full text] [doi: 10.3389/fdgth.2021.686720] [Medline: 34713157]
36. Ordun C, Purushotham S, Raff E. Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. arXiv Preprint posted online May 6, 2020.
37. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. J Med Internet Res 2020 Apr 28;22(4):e19118 [FREE Full text] [doi: 10.2196/19118] [Medline: 32302966]
38. Debnath R, Bardhan R. India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling. PLoS One 2020;15(9):e0238972 [FREE Full text] [doi: 10.1371/journal.pone.0238972] [Medline: 32915899]
39. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A 2004 Apr 06;101 Suppl 1:5228-5235 [FREE Full text] [doi: 10.1073/pnas.0307752101] [Medline: 14872004]
40. Huang Y, Lowe HJ, Klein D, Cucina RJ. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. J Am Med Inform Assoc 2005;12(3):275-285 [FREE Full text] [doi: 10.1197/jamia.M1695] [Medline: 15684131]
41. Creswell JW, Creswell JD. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 5th ed. Thousand Oaks, CA: Sage Publications; 2018.
42. Miles M, Huberman A, Saldana J. Qualitative Data Analysis: A Methods Sourcebook (3rd edition). Thousand Oaks, CA: Sage Publications; 2013.
43. Mohammad S, Turney P. Crowdsourcing a Word-Emotion Association Lexicon. Comput Intell 2013;29:436 [FREE Full text] [doi: 10.1111/j.1467-8640.2012.00460.x]
44. FÅ N. Evaluation of a word list for sentiment analysis in microblogs. arXiv Preprint posted online March 15, 2011.
45. Bing L. Sentiment Analysis and Subjectivity. In: Handbook of Natural Language Processing (2nd edition). London: Chapman and Hall/CRC; 2010.
46. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 2011 Jun;37(2):267-307 [FREE Full text] [doi: 10.1162/coli_a_00049]
47. Brooke J, Tofiloski M, Taboada M. Cross-Linguistic Sentiment Analysis: From English to Spanish. 2009 Presented at: International Conference RANLP 2009; 2009; Borovets p. 50-54.
48. Hu M, Liu B. Mining Opinion Features in Customer Reviews. 2004 Presented at: Nineteenth National Conference on Artificial Intelligence; July 25-29, 2004; San Jose, CA.
49. Vilares D, Alonso PM, Gómez-Rodríguez C. Polarity classification of opinionated Spanish texts using dependency parsing. Procesamiento de Lenguaje Natural 2013;50:13-20.
50. Spasic I, Nenadic G. Clinical Text Data in Machine Learning: Systematic Review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
51. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence. J Med Internet Res 2020 Aug 18;22(8):e22590 [FREE Full text] [doi: 10.2196/22590] [Medline: 32750001]
52. Sahoo S, Mehra A, Dua D, Suri V, Malhotra P, Yaddanapudi LN, et al. Psychological experience of patients admitted with SARS-CoV-2 infection. Asian J Psychiatr 2020 Dec;54:102355 [FREE Full text] [doi: 10.1016/j.ajp.2020.102355] [Medline: 33271684]
53. Shirai M, Soshi T. Why is heartache associated with sadness? Sadness is represented by specific physical pain through verbal knowledge. PLoS One 2019;14(5):e0216331 [FREE Full text] [doi: 10.1371/journal.pone.0216331] [Medline: 31042783]
54. Wu DJ, Svoboda RC, Bae KK, Haase CM. Individual differences in sadness coherence: Associations with dispositional affect and age. Emotion 2021 Apr;21(3):465-477. [doi: 10.1037/emo0000731] [Medline: 32191094]
55. Gautret P, Million M, Jarrot P, Camoin-Jau L, Colson P, Fenollar F, et al. Natural history of COVID-19 and therapeutic options. Expert Rev Clin Immunol 2020 Dec;16(12):1159-1184. [doi: 10.1080/1744666X.2021.1847640] [Medline: 33356661]
56. Hodgens A, Sharman T. Corticosteroids. Treasure Island, FL: StatPearls Publishing; 2022.

57. Pancorbo-Hidalgo P, García-Fernández F, Pérez-López C, Soldevilla AJ. Prevalence of pressure injuries and other dependence-related skin lesions in adult patients admitted to Spanish hospitals: the fifth national study in 2017. Gerokomos 2019;30(2):86 [FREE Full text]
58. Alderden J, Rondinelli J, Pepper G, Cummins M, Whitney J. Risk factors for pressure injuries among critical care patients: A systematic review. Int J Nurs Stud 2017 Jun;71:97-114 [FREE Full text] [doi: 10.1016/j.ijnurstu.2017.03.012] [Medline: 28384533]
59. Ruiz-García A, Arranz-Martínez E, García-Álvarez JC, García-Fernández ME, Palacios-Martínez D, Montero-Costa A, En representación del Grupo de Investigación del Estudio SIMETAP. Grupo de Investigación del Estudio SIMETAP:. Prevalence of diabetes mellitus in Spanish primary care setting and its association with cardiovascular risk factors and cardiovascular diseases. SIMETAP-DM study. Clin Investig Arterioscler 2020;32(1):15-26. [doi: 10.1016/j.arteri.2019.03.006] [Medline: 31130360]
60. Delpont B, Blanc C, Osseby G, Hervieu-Bègue M, Giroud M, Béjot Y. Pain after stroke: A review. Rev Neurol (Paris) 2018 Dec;174(10):671-674. [doi: 10.1016/j.neurol.2017.11.011] [Medline: 30054011]
61. Chang HM, Chiou SF, Liu HY, Yu HC. Using a Text-Mining Approach to Evaluate the Quality of Nursing Records. Stud Health Technol Inform 2016;225:813-814. [Medline: 27332355]
62. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 01;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
63. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. J Med Internet Res 2021 Feb 10;23(2):e25431 [FREE Full text] [doi: 10.2196/25431] [Medline: 33497352]

Abbreviations

AUC: area under the curve
EHR: electronic health record
ENCN: electronic nursing clinical notes
LDA: Latent Dirichlet Allocation
NRC: National Research Council of Canada
STM: structural analysis of thematic models
SVM: Support Vector Machine

Edited by C Lovis; submitted 28.03.22; peer-reviewed by JF Velarde-García, P Parás-Bravo; comments to author 11.04.22; revised version received 12.04.22; accepted 21.04.22; published 12.05.22.

Please cite as:
Cuenca-Zaldívar JN, Torrente-Regidor M, Martín-Losada L, Fernández-De-Las-Peñas C, Florencio LL, Sousa PA, Palacios-Ceña D
Exploring Sentiment and Care Management of Hospitalized Patients During the First Wave of the COVID-19 Pandemic Using Electronic Nursing Health Records: Descriptive Study
JMIR Med Inform 2022;10(5):e38308
URL: https://medinform.jmir.org/2022/5/e38308
doi:10.2196/38308
PMID:354869

Original Paper

Deep Neural Networks for Simultaneously Capturing Public Topics and Sentiments During a Pandemic: Application on a COVID-19 Tweet Data Set

Adrien Boukobza[1,2,3]; Anita Burgun[1,2,3], MD, PhD; Bertrand Roudier[4], PharmD, PhD; Rosy Tsopra[1,2,3], MD, PhD

[1]Université Paris Cité, Sorbonne Université, Inserm, Centre de Recherche des Cordeliers, Paris, France

[2]Inria, HeKA, PariSanté Campus, Paris, France

[3]Department of Medical Informatics, Assistance Publique – Hôpitaux de Paris, Hôpital Européen Georges-Pompidou, Paris, France

[4]ESIEE, Cité Descartes, Noisy le Grand Cedex, France

Corresponding Author:
Adrien Boukobza
Department of Medical Informatics
Assistance Publique – Hôpitaux de Paris
Hôpital Européen Georges-Pompidou
20 rue Leblanc
Paris, F-75015
France
Phone: 1 156092167
Email: hadrien_b@hotmail.fr

Abstract

Background: Public engagement is a key element for mitigating pandemics, and a good understanding of public opinion could help to encourage the successful adoption of public health measures by the population. In past years, deep learning has been increasingly applied to the analysis of text from social networks. However, most of the developed approaches can only capture topics or sentiments alone but not both together.

Objective: Here, we aimed to develop a new approach, based on deep neural networks, for simultaneously capturing public topics and sentiments and applied it to tweets sent just after the announcement of the COVID-19 pandemic by the World Health Organization (WHO).

Methods: A total of 1,386,496 tweets were collected, preprocessed, and split with a ratio of 80:20 into training and validation sets, respectively. We combined lexicons and convolutional neural networks to improve sentiment prediction. The trained model achieved an overall accuracy of 81% and a precision of 82% and was able to capture simultaneously the weighted words associated with a predicted sentiment intensity score. These outputs were then visualized via an interactive and customizable web interface based on a word cloud representation. Using word cloud analysis, we captured the main topics for extreme positive and negative sentiment intensity scores.

Results: In reaction to the announcement of the pandemic by the WHO, 6 negative and 5 positive topics were discussed on Twitter. Twitter users seemed to be worried about the international situation, economic consequences, and medical situation. Conversely, they seemed to be satisfied with the commitment of medical and social workers and with the collaboration between people.

Conclusions: We propose a new method based on deep neural networks for simultaneously extracting public topics and sentiments from tweets. This method could be helpful for monitoring public opinion during crises such as pandemics.

(JMIR Med Inform 2022;10(5):e34306) doi:10.2196/34306

KEYWORDS

neural network; deep learning; COVID-19; explainable artificial intelligence; decision support; natural language processing

Introduction

Background

Pandemics caused by emerging pathogens are public health emergencies. They have dramatic consequences for the population (mortality, morbidity, social life) and the economy [1]. The number of outbreaks has increased in recent decades, and this trend is expected to intensify [1] in the next years. In particular, when the first cases of pneumonia caused by the SARS-CoV-2 pathogen were declared in Wuhan, Hubei Province, China [2,3], the virus rapidly spread around the world, leading the World Health Organization (WHO) to declare a pandemic on March 11, 2020, and announced it on Twitter with the tweet: "BREAKING "We have therefore made the assessment that #COVID19 can be characterized as a pandemic"-@DrTedros #coronavirus." With this declaration occurring on social media, Twitter remains an ideal medium to study public opinion on the declaration of the COVID pandemic.

Utility of Social Networks for Identifying Sentiments and Topics of the Population During Pandemics

As public engagement is a key element for mitigating pandemics [4-6], several studies have already mined social media since the beginning of the COVID-19 pandemic but with distinct objectives (eg, infoveillance) [7-9] or during different periods (eg, when first important measures were taken in the United States) [7,10-14]. To our knowledge, there is no study analyzing public opinion in the immediate reaction just after the WHO announcement.

Social networks have largely been used to capture public opinion, especially during outbreaks (eg, Ebola [15], H1N1 [16]). The methods used to analyze texts from social networks have considerably improved over time: manual analysis first, followed by natural language processing (NLP) approaches based on syntactic-semantic or statistical techniques [17], and more recently, deep learning approaches [18,19]. Deep learning methods provide new perspectives on text analysis since they give the possibility to (1) integrate semantic information around text (eg, with pretrained word embedding, which allows higher semantic information as the input for the neural network rather than a one-hot encoder [20]) and (2) analyze a significantly larger corpus of text nearly in real time, making it possible to discover new evidence faster [21]. These approaches [7,8,22-26] have already been used to capture topics (eg, for the Covid Infoveillance study [7] or insulin pricing concerns in the United States [27]) or sentiments (eg, on social network posts or on health care tweets [17,28-30]).

Prior Work With Topic Extraction

Several approaches have been used for topic extraction, including qualitative analysis, descriptive analysis, and topic analysis.

Qualitative Analysis

Qualitative analyses [22,23,31] capture common themes from manual analysis, fragmentation, and labelling of text. This method has demonstrated its capacity to accurately capture new and complex topics [32] but with some major issues: It requires human coders, time, and resource consumption and is not suitable for use with high-dimensional data.

Descriptive Analysis

Descriptive analyses [8] capture the distribution of word frequencies by studying the repetition of words among topics identified from the internet. It allows researchers to correlate the importance of a topic to the volume of searches among this peculiar topic. The main pitfall of this method is the inability to consider the context around the word.

Topic Analysis

Topic analysis is a method used to discover topics that occur in a collection of documents and has largely been used to mine social media. This method aims at identifying patterns in documents using NLP approaches. Two main categories of topic analyses are commonly used: topic classification [33] and topic modeling [34].

Topic classification uses supervised learning algorithms (eg, Naïve Bayes [19], support vector machine [SVM] [35]) that need to be trained beforehand with labeled documents, consequently requiring a priori knowledge of corpus topics. These algorithms can achieve variable performance, with a precision varying from 44.9% to 93.3% [19], depending on the methods used.

On the contrary, topic modeling uses unsupervised learning algorithms that do not need to be trained beforehand. They are thus less work-intensive than supervised learning algorithms since they do not need human-labelled data but often require larger data sets and are less precise than supervised learning algorithms. Latent semantic analysis is the traditional method for topic modeling [36]. It is based on the distributional hypothesis and assumes that words with close meaning will occur in similar pieces of text [37]. This assumption enabled the development of algorithms such as latent Dirichlet allocation (LDA) [7,25,26,38], which is popular in the medical domain [39]. This algorithm identifies latent topics from words tending to occur together and outputs *n* clusters of words grouped together by similarity. The topics are then manually labelled according to the interpretation of the set of words within each cluster [7,40]. However, LDA requires the investigator to predefine the number of topics and does not consider the sequence of words [39]. Topic modeling has been poorly assessed, perhaps a result of the difficulty comparing the clusters obtained with a gold standard. To overcome this lack of evidence, Zhang et al [38] proposed an original approach for assessing LDA: They compared the topics extracted from LDA to those collected through a national questionnaire survey and reported a kappa concordance coefficient of 0.72.

Prior Work With Sentiment Analysis

Several approaches have been used for sentiment analysis, including lexicon-based methods, supervised machine learning methods, and hybrid methods.

Lexicon-Based Methods

Lexicon-based methods are unsupervised methods that do not require training an algorithm and depend only on existing dictionaries [29]. These methods assume that the polarity of a

text (positive or negative) can be obtained by characterizing the constituent words within [29]. A key argument for their adoption was the fact that they only compute the number of positive and negative words [41] and thus are faster to implement. They are also easily adaptable to various languages by using language-specific dictionaries [42]. However, they present some limitations that come with language analysis, especially regarding negation, sarcasm, or words with different meaning [28,29]. Furthermore, they are essentially limited by the size, coverage, and quality of the dictionary [17]. Interestingly, lexicon-based methods can achieve an accuracy up to 94.6% [43], depending on the dictionary used [43-46].

Supervised Machine Learning Methods

Supervised machine learning methods, which require time to be trained, have also been used [47]. Naïve Bayes often better operates on well-shaped data, whereas SVM often achieves better results with low-shaped data. As social media are poor-quality data, due to very varying length of tweets, colloquial language, and numerous spelling mistakes, larger training data sets are needed to achieve good performance, and the complexity of these methods may impact training time [48]. They can achieve variable performance, with reported accuracies ranging from 48% to 91% [47,49,50], depending on the algorithm used.

Hybrid Approaches

Hybrid approaches combine both previous methods. In a recent literature review, Drus and Khalid [29] demonstrated that hybridized approaches to sentiment analysis often outperform lexicon-based or machine learning–based approaches alone. For example, Hassan et al [47] used lexicon annotation and multinomial Naïve Bayes for depression measurement from social networks and reported an accuracy rate of 91%; Zhang et al [51] used lexicon annotation and SVM to annotate sentiments from tweets and reported an accuracy of 85.4%.

Prior Work Aiming to Capture Both Topics and Sentiments

Few methods based on topic-sentiment models have been developed, including the joint sentiment topic (JST) model, Topic-Sentiment Mixture (TSM) model, and Time-aware Topic Sentiment (TTTS) model.

Joint Sentiment Topic Model

The JST [52] model is a probabilistic modelling framework that extends LDA with a new sentiment layer. JST is fully unsupervised and extracts both topics and sentiments at a document level [52]. However, JST ignores the word ordering (bigrams or trigrams [52]). Reverse JST [53] is derived from JST with an inversion of the order of the topic and sentiment layers. The Aspect and Sentiment Unification Model (ASUM) [54] is close to JST but focuses on the sentence level. These models have been poorly assessed and were essentially applied on nonmedical data sets, with an accuracy varying from 59.8% to 84.9% for JST [52,53] and 69.5% to 75.0% for reverse JST [53].

Topic-Sentiment Mixture Model

TSM [55] is based on the probabilistic latent semantic indexing model and includes an extra background component and 2 sentiment subtopics. It has been assessed on various weblog data sets [55] but suffers from problems of inferencing on new documents and overfitting data [52] and requires postprocessing to obtain the sentiment [56].

Time-Aware Topic Sentiment Model

More recently, the TTTS model [57] is a joint model for topic-sentiment evolution, based on LDA and allowing analysis of topic-sentiment evolution over time [57].

Strengths and Weaknesses of Previous Work

Many approaches have proven useful for identifying public topics alone but without the associated sentiment. Other works, especially hybrid approaches, have proven useful for sentiment detection alone but cannot capture the topics alongside sentiment detection.

In both cases, this makes the results less informative and useful [52]. Simultaneously capturing topics and sentiments would be more relevant for better comprehension of public opinion [52], especially in a time of crisis. Topic-sentiment models have been proposed for the simultaneous capture of public opinion and sentiments but may require prior domain knowledge and have not been applied yet to the medical and social media domains [52,53,55].

Potential for a Neural Network–Based Approach to Advance This Area of Research

Neural networks have achieved impressive performances in many NLP tasks, such as sentiment prediction [58-60]. Furthermore, the probabilities generated by neural networks could be used to represent sentiment intensity through a quantitative scale leading to more precise information than basic sentiment classification into dual qualitative classes (negative or positive). Surprisingly, to our knowledge, they have not been used yet for the simultaneous capture of public topics and sentiments from social media.

Here, we propose incorporating convolutional neural networks (CNNs) in conjunction with sentiment lexica to simultaneously capture public topics and sentiments in a hybridized approach [18,29]. The simultaneous capture of public topics and sentiments, without prior knowledge, would be very useful during crises, such as the COVID-19 outbreak.

Methods

Preparation of the Tweet Data Set for Use as an Input for Neural Networks

Data Collection

To analyze the immediate effect of the announcement of the COVID-19 pandemic by the WHO, we focused on tweets relating to coronavirus posted on Twitter the day after the announcement. We collected all tweets containing the keywords “coronavirus” or “COVID” posted in English as recognized by Twitter services on March 12, 2020 (ie, from 00:00:01 to

23:59:59). For each tweet, we extracted the tweet ID, text content, and time stamp. We also filtered them using the language parameter of Twint Python Library [61] to allow the extraction of English-written tweets only. We verified the absence of tweets in other language by using common stop words of these languages, resulting in only finding foreign city names or family names.

We extracted 1,386,496 tweets from Twitter's database with the Twint Python library and stored them in the JSON format.

Ethical Approval

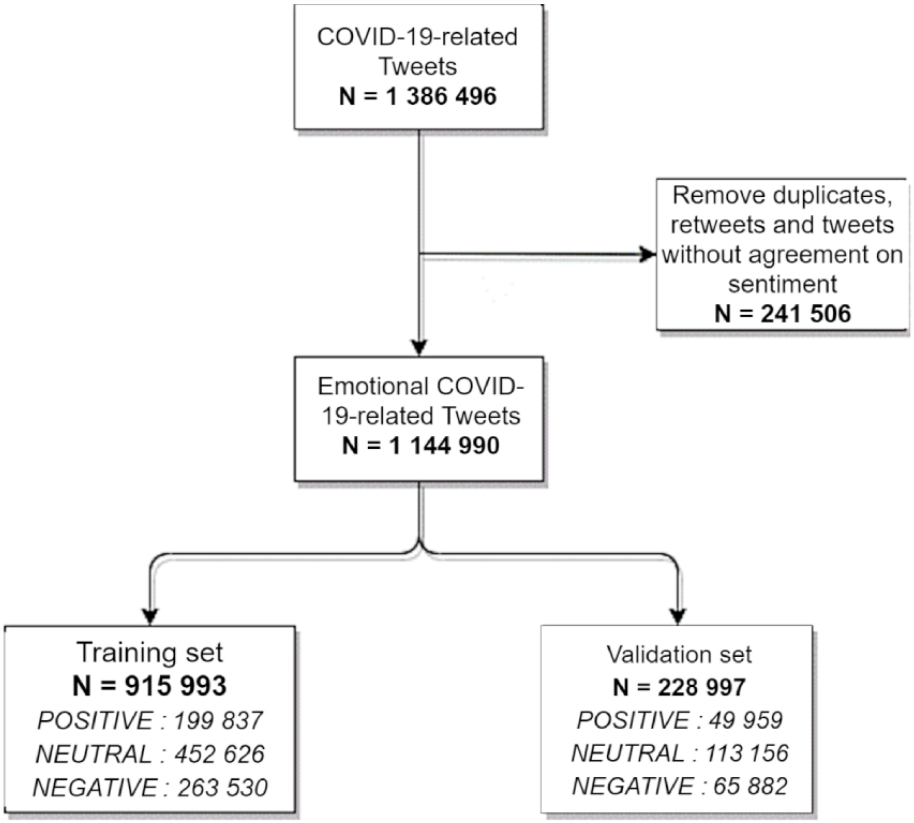
Ethic approval was not needed as analysis of large bodies of text written by humans on the internet and in some social media such as Twitter (eg, quantitative analysis such as infodemiology or infoveillance studies or for qualitative analysis) is not considered "human subjects research."

Data Preprocessing

We removed 241,506 (17.5%) duplicate tweets and retweets to limit the risk of overrepresentation of one person's view. Twitter elements (URLs, links to pictures, hashtags, mentions), punctuation, isolated letters, and typographic UTF-8 characters, such as stylized commas or apostrophes, were also removed. Likewise, stop words from Porter's list [62] were removed using the Python library Natural Language Toolkit (NLTK) [63], with orthographic variations. Tweet content was then lower-cased, and "coronavirus" and "COVID" were mapped under a unique term.

Figure 1 provides a flow chart of tweet collection, preprocessing, and splitting into the training and testing sets.

Figure 1. Study flowchart.



Sentiment Annotation

Each tweet was automatically annotated with 3 sentiment labels from 3 different sentiment lexicons from R package tidytext [64] (AFINN [44], BING [43], and NRC [45,46]). These lexicons have largely been used in previous works [30,42,44]. Each lexicon provided a numerical value for each sentiment word in the tweet, and these values were summed to annotate the general sentiment of the tweet for each lexicon considered, as described in other works [41,42]. Thus, for each annotation, the sum value could be positive, equal to 0, or negative resulting in positive, neutral, or negative annotation by the considered sentiment lexicon.

Annotation conflicts were handled using a simple rule-based algorithm to compute a single annotation for each tweet. This algorithm is based on the majority vote method and produced a unique qualitative annotation as "positive," "neutral," or "negative." If a majority vote was not obtained (ie, if each algorithm returned a different statement), the tweets were excluded from the data set.

The automatic annotation of included tweets was controlled on 50 randomized tweets, using a manual revision of tweet annotation, resulting in an overall agreement of 86% between algorithm and manual annotation, resulting in a kappa coefficient score of 0.73.

Deep Neural Networks for Simultaneously Capturing Public Topics and Sentiments

Tokenization, Word Embedding, and CNN Architecture

CNN architecture was chosen as it is known to consider Ngrams, making various levels of analysis possible.

All words in each tweet were tokenized, and tweets were postpadded for use as input into the pretrained embedding layer of the neural networks, which encoded semantic properties for each token. We used a 25-dimension Global Vector for word representation (GloVe) embedding trained on 2 billion tweets to shorten training time and achieve better results. This embedding is available from the GloVe project page [65].

The resulting vectors were then passed to a convolutional unit composed of a convolutional layer (able to analyze unigrams, bigrams, or trigrams), global max pooling layer, dense layer, and dropout layer for regularization and prevention of overfitting. A final dense layer composed of 3 units alongside a softmax activation function computed the probabilities of the tweet belonging to each class of sentiment (positive, neutral, negative). Early stopping was used to prevent overfitting when training our models.

To perform the supervised learning step, the data set was split using stratification over sentiment annotation, allocating 80% (915,993 tweets) for training and 20% for validation (228,997 tweets; Figure 1). The best model was found after 10 training iterations and used a kernel size of 2 on the convolutional layer. The accuracy was 81%, and the F1 score was 81% on the validation data set (Table 1).

Table 1. Performance of the neural network for sentiment prediction.

Performance measure	Positive	Neutral	Negative	Total
Accuracy	83%	80%	82%	81%
F1 score	79%	82%	81%	81%
Precision	77%	85%	79%	82%
Recall	82%	80%	83%	81%

Neural Network Outputs: Sentiment Intensity Score and Weighted Word Capture

For each tweet, we captured the dominant sentiment as a sentiment intensity score that was calculated from the 3 probabilities predicted by the CNN:

SIS = P(POSITIVE) x 1 + P(NEUTRAL) x 0 + P(NEGATIVE) x (–1)

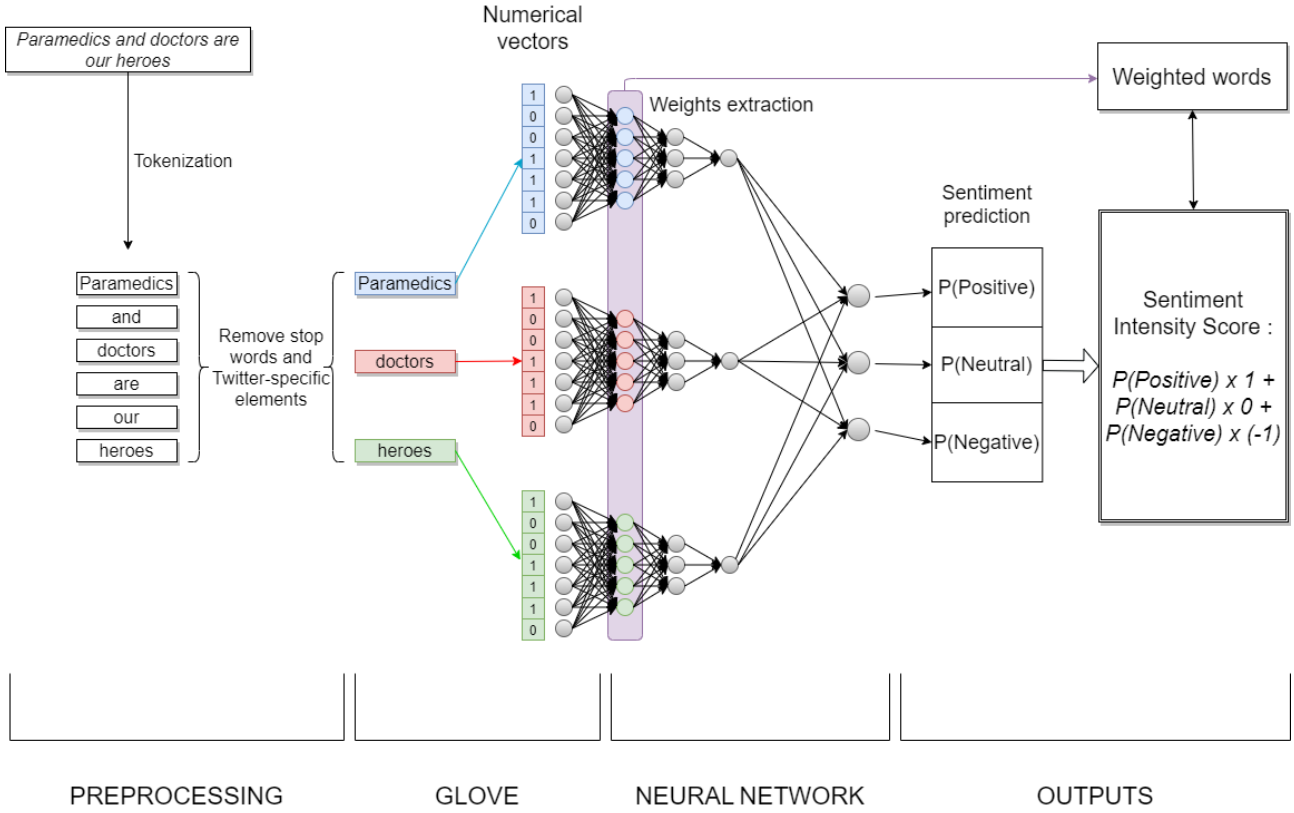
where SIS, P(POSITIVE), P(NEUTRAL), and P(NEGATIVE) are sentiment intensity score and probabilities for a tweet to belong to the positive, neutral, and negative sentiment classes, respectively, according to the neural network.

Applying this formula allowed us to distinguish 21.82% (249,796/1,144,990) of the tweets as positive, 49.41% (565,782/1,144,990) as neutral, and 28.77% (329,412/1,144,990) as negative. The sentiment intensity score of each tweet was then represented on a scale from –100% (totally negative) to +100% (totally positive), permitted by using the softmax activation function.

As the CNN architecture alternates convolutional and pooling layers, it allows, first, aggregation of the numerical input coming from each word separately until a hidden layer and then combination of the values of this hidden layer until the output of the CNN. Hence, this hidden layer encompasses a value for each word, and this value can be seen as a contribution score (or a weight) of each word in the computation of the final output of the CNN [66]. As the output of the CNN is used to compute the dominant sentiment intensity of the whole tweet, the intermediate values extracted from the hidden layers make it possible to associate “weighted words” to the sentiment intensity score of the tweet. Figure 2 summarizes the capture of the sentiment intensity score and of the weighted words.

In previous steps, the weighted words and sentiment intensity score were captured at the individual tweet level. At the tweet data set level, we computed the average weight of each word for each sentiment intensity score by gathering similar words from distinct tweets and applying a mean function. The resulting matrix contained the weighted words for each given sentiment intensity score.

Figure 2. Neural network outputs, where P(POSITIVE), P(NEUTRAL), and P(NEGATIVE) are the probabilities for a tweet to belong to the positive, neutral, and negative sentiment classes, respectively, according to the neural network. Please note that the convolutional neural network (CNN) is represented here as a simple perceptron to facilitate reading, and each word's contribution score is represented with colored neurons.



Visualization of Neural Network Outputs

We developed a Shiny [67] application (available at [68]) based on word cloud representation to visualize the weighted words for each sentiment intensity score. This application provides 2 panels: On the right panel, the word cloud displays the weighted words for a given sentiment intensity score. On the left panel, the word cloud can be customized through options specifying the sentiment intensity score, the number and type of words to display (coronavirus or sentiment-related terms), and the esthetics (eg, palette of colors, total percentage of vertical words, and use of a radial gradient).

To generate our word clouds, we replaced the use of word frequencies to summarize text documents by the weights calculated in our matrix. The visualization was made clearer by grouping all lexical variants of a word together, using the word lemmatizer from the R package textstem [69]. We also implemented options allowing the user to ignore all sentiment words and emojis, to choose the word count threshold for display, and to choose the precision of the sentiment score (integer or float to 1 or 2 decimal places).

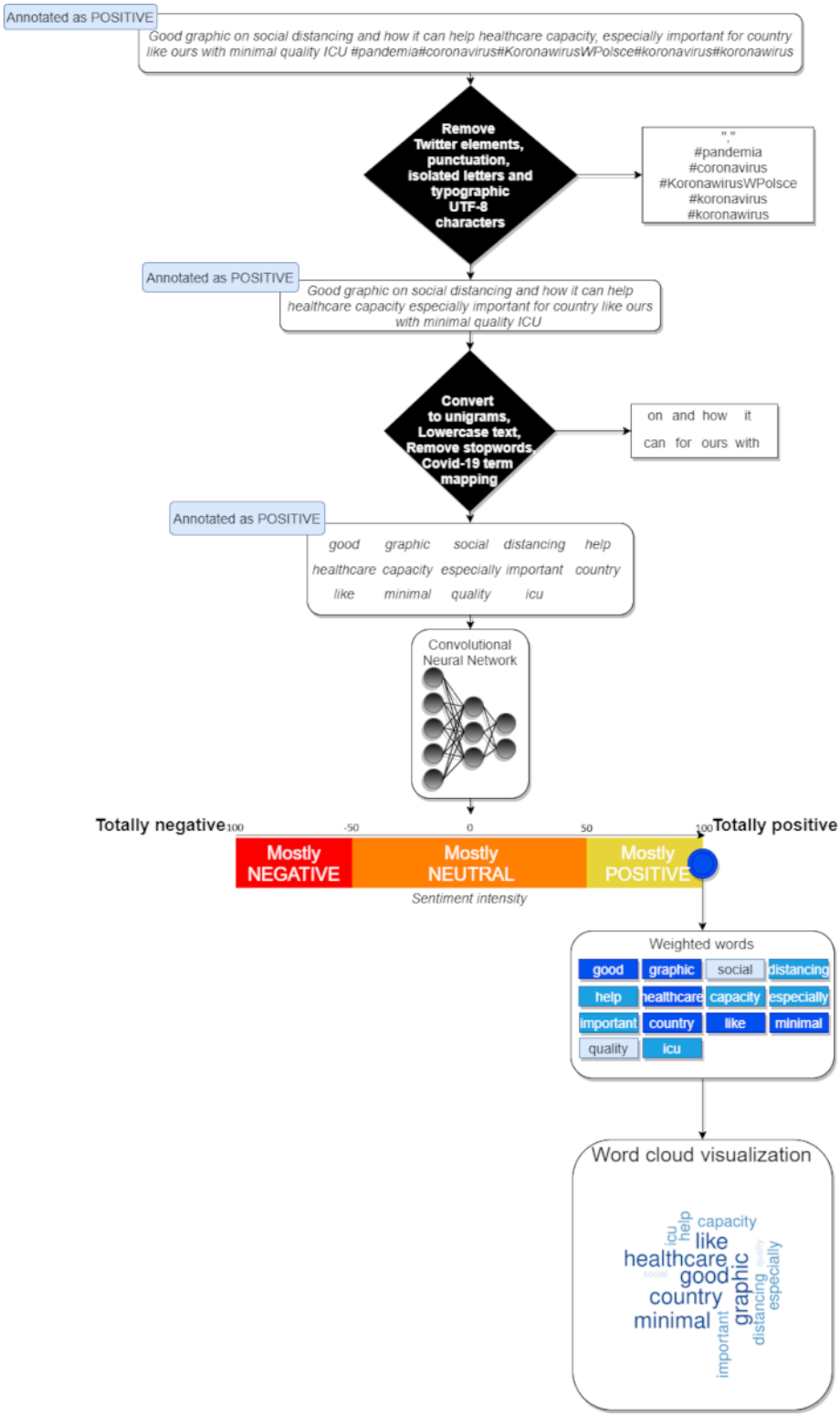
Identification of the Main Topics Discussed by the Public and Their Associated Sentiment Intensity

Using the Shiny interface, we captured the highest weighted words for the most extreme sentiment intensity scores (negative sentiment: –100; positive sentiment: +100). Author A Boukobza then manually analyzed the top 100 words for both extreme sentiments using string-matching techniques and identified main negative and positive topics within tweets. Each topic was assigned by the manual analysis of these words. Then, we calculated the number of tweets discussing each topic within the data set.

In the results section, we replaced the real names of politicians, political parties, websites, and media with anonymous epithets such as "politicianX," "politicalPartyX," "webX," "mediaX."

Figure 3 summarizes the general method used for extracting weighted words and their associated sentiments from Twitter data.

Figure 3. Method used for simultaneously extracting weighted words and their associated sentiments from tweets. An example of a tweet at each step is provided, from initial preprocessing to sentiment intensity scale classification (here, the tweet sentiment score is +100%) and final output as a word cloud.



Results

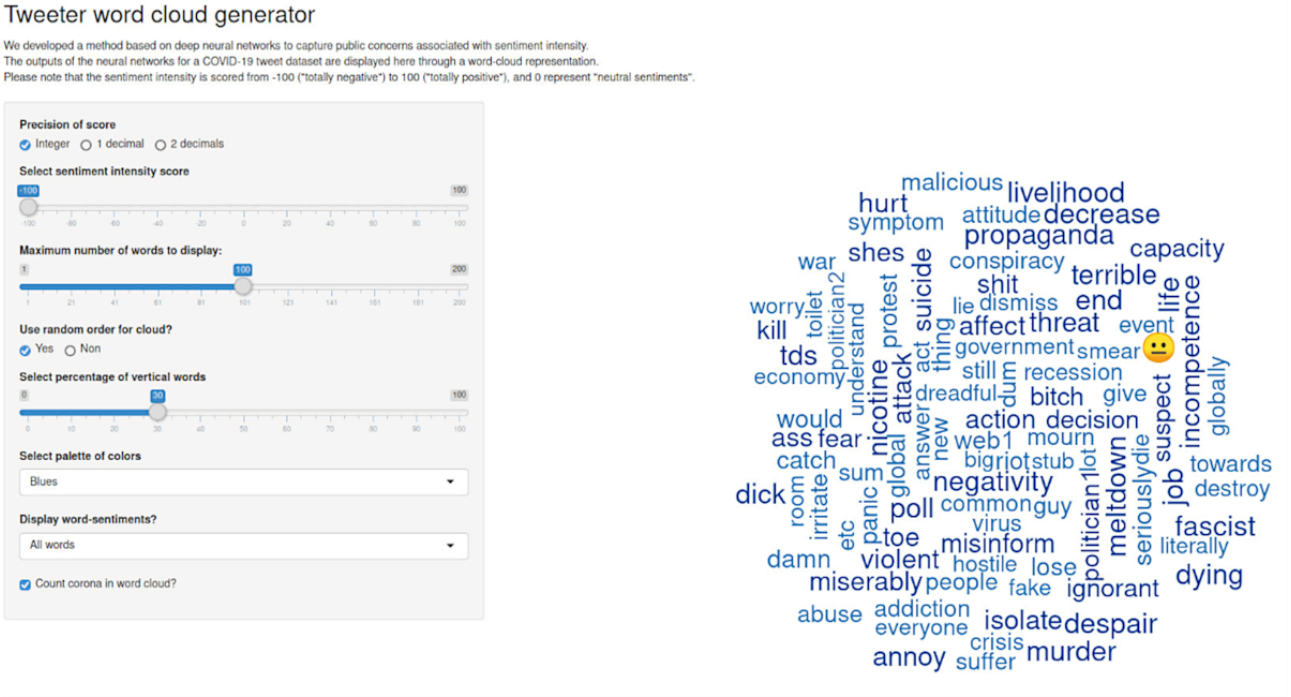
Visualization of Neural Network Outputs With an Interactive Interface

Neural network outputs were visualized with an interactive interface displaying a word cloud composed of the weighted words for each sentiment intensity score.

The analysis of the top 100 most important words for each class allowed us to predistinguish main themes retrieved for positive, negative, and neutral classes. In the totally positive class (ie, +100 sentiment intensity score), the top 100 words included words such as “happiness,” “democratic,” “ethical,” “quarantine,” or “expertise.” Concerning the neutral class (ie, 0 sentiment intensity score), the top 100 words included names

(eg, “François,” “Eliott”), adverbs (eg, “thankfully,” “formally”), or scientific words (eg, “petri,” “aneurysm”). In the totally negative class (ie, –100 sentiment intensity score), the top 100 words included words such as “job,” “economy,” “afraid,” “panic” (Figure 4).

Figure 4. Interactive web application for visualizing neural network outputs. The real names of politicians, political parties, websites, and media were replaced by anonymous epithets such as “politicianX,” “politicalPartyX,” “webX,” “mediaX.”.



Identification of Public Topics and Associated Sentiment Intensity

Using word cloud analysis, we captured the topics for both extreme positive and negative sentiment intensity scores that were discussed in Twitter in immediate reaction to the announcement of the pandemic by the WHO. The analysis of these topics revealed that public opinion was extremely negative about the consequences of the pandemic on the economy and health care system. Conversely, public opinion was extremely positive regarding the mutual aid and cooperation between people and the public health measures taken against the spread of COVID-19. More details are given in the following sections, and example tweets are provided in Table 2.

Table 2. Main positive and negative topics, with highest weighted words, illustrative tweets, and the number of tweets containing the weighted word.

ID	Topics	Weighted words identified by the neural network	Example of an original tweet	Number of tweets containing weighted words for each topic
Negative topics				
1	International situation	italian, china, eu, euro, italy, politican1, politicalParty1, politicalParty2, politician2, president, government, politician3, incompetence, fascist	*Italy* is already today worse affected by Covid-19 than *China*. (...)	11,297
2	Economy	job, impact, industry, yougov, hire, financial, market, livelihood, diarrhea, recession, economy	(...) The *markets* are trash, every *industry* is freaking out, and people are losing their *jobs* because it's stalling the *economy* and no one is *hiring*. (...)	3486
3	Media and social media	media1, media2, american	Signed out of my *media1*. (...) *Media2* is a HORRIBLE thing to be on with this damn Coronavirus (...)	1428
4	Media and social media	media1, media2, american	Can the media be declared enemies of the people? They (...) lie to us, (...) and fail to report news/statistics that *Americans* need to know. (...)	
5	Medical situation	ventilator, paramedic, triage, ration, supply	The (...) most dreadful thing we might face is *rationing* or *triaging* who gets *ventilators*. Emergency rooms across the U.S (...) have limited capacity and *supplies* (...)	1411
6	Public health measures	stay, senior, travel, indoor, cancel, ban	The EU *travelban* (...) I must admit is terrible decision extremely terrible (...)	8396
7	COVID-19 origin	coronavirusHoax, fake, conspiracy, propaganda	(...) the *Fake* News Media are fabricating the hype and panic to destroy the economy (...) #Pandumbic *#coronavirusHoax*	1680
Positive topics				
8	International situation	italy, nhs, democracy, gov, politician4	(...) Freer and more *democracy* countries can do this if they take needed measures.	2178
9	Economy	client, colleague, customer, company	We would like to extend our heartfelt appreciation to all of our *clients* and partners working on the frontlines (...)	745
10	Medical situation	mask, research, health, healthy, resources, healthcare, doctor, applause, hero	Put all your money and *resources* into getting the cure for the Coronavirus you look like a hero and win the election	4803
11	Public health measures	stay, control, announce, interpersonal, family, canceleverything, relative, country, precaution, sanitation, icu, measures, prevention, protect	Good graphic on social distancing and how it can help healthcare capacity, especially important for a *country* like ours with minimal quality ICU #pandemia #coronavirus #KoronawirusWPolsce #koronavirus #koronawirus	6642
12	Mutual aid and cooperation	collaborative, together	(...) Communities who work *together* to ensure the health and well-being of their fellow neighbor will be stronger and healthier than those who don't. #Coronavirus	470

The 6 Main Negative Public Topics Discussed on Twitter in Immediate Reaction to the Announcement of the Pandemic by the WHO

Regarding the international situation, Twitter users were worried about the situation in Italy (eg, the number of cases exceeding those in China; Table 2, ID 1) or the risk of punishment or imprisonment for Italians not respecting lockdown. They also discussed travel bans and their consequences, such as the US decision to ban all flights to Europe at a time at which only Italy had a major COVID-19 epidemic. Crisis management and decisions taken by politicians, such as decisions relating to paramedical staff management, were also highly criticized. Regarding economy, Twitter users expressed their fears about the economic consequences of COVID-19. They were worried about the shortages induced by panic buying, such as those leading to a shortage of toilet rolls, and anxiety about the possibility of losing their jobs and being unable to pay their debts (Table 2, ID 2). They also mentioned a potential global recession crisis, caused partly by flight limitations. Regarding media and social media, Twitter users were angry with the media and social media, which they blamed for amplifying fears and stress relating to COVID-19 (Table 2, ID 3), and for not reporting COVID-19 statistics (Table 2, ID 4). Regarding the

medical situation, Twitter users were concerned about the medical situation, particularly the management of paramedical staff and materials. They expressed worries about the small number of ventilators available and the likely consequences in terms of equality of access to health care (Table 2, ID 5). Regarding public health measures, Twitter users complained about the limitations of personal liberties, such as the prohibition of flights to Europe (Table 2, ID 6) and the canceling of many events. Regarding the COVID-19 origin, Twitter users talked about "CoronavirusHoax." They suggested that the pandemic was a hoax and that COVID-19 was a fake disease and evoked a conspiracy theory driven by economic and political motives (Table 2, ID 7).

The 5 Main Positive Public Topics Discussed on Twitter in Immediate Reaction to the Announcement of the Pandemic by the WHO

Regarding the international situation, Twitter users expressed their satisfaction with the actions and decisions taken by some countries, such as Japan, Hong Kong, Singapore, South Korea (Table 2, ID 8), or Denmark (eg, the decision to impose a lockdown at the right timing). They also highlighted the efficient measures taken by some countries such as the United Kingdom to overcome the negative effects of lockdown (eg, National Health Service access or online courses for students). Regarding the economy, Twitter users were very grateful to all those who worked during the crisis (Table 2, ID 9). Public workers were even described as "people working hard for ensuring population security." Twitter users were also informed about the continuity of services ensured by some private companies despite the crisis. They were satisfied with the health measures taken by these companies (eg, social distancing, sanitizing measures, provision of masks). Regarding the medical situation, Twitter users maintained their trust and hope regarding the medical situation. They highly appreciated the work of medical and paramedical staff and their involvement in communicating reliable information about COVID-19 to the population. They highlighted the importance of developing telemedicine and evoked the possibility of a COVID-19 vaccine and its potential consequences for health policies (Table 2, ID 10). They also discussed the production and free distribution of infographics and masks to health professionals by private companies. Regarding public health measures, Twitter users encouraged the respect of national measures, social distancing, and lockdowns to allow people to protect themselves and their families. They also appreciated the graphics providing guidance on the changes in behavior required to limit the spread of coronavirus (Table 2, ID 11). Regarding mutual aid and cooperation, Twitter users were satisfied with the level of cooperation between people in front of the coronavirus crisis (Table 2, ID 12). They were grateful to workers and medical and paramedical staff.

Discussion

Principal Findings

We proposed here an original new approach based on deep neural networks for the simultaneous capture of public topics and sentiments from Twitter data. We trained a CNN on a training data set of 915,993 tweets and achieved a performance of 81% for both accuracy and F1 score. The trained neural network was able to capture the weighted words and their associated sentiment intensity score. These outputs were then visualized through an interactive and customizable web interface displaying the weighted words as a word cloud representation. The trained model was then used to analyze public topics and sentiments in reaction to the announcement of the COVID-19 pandemic by the WHO.

Strengths and Limitations

Our study has several strengths. We combined lexicons and deep learning approaches to improve sentiment prediction. We used CNN to capture simultaneously weighted words associated with sentiment intensity score and to compare unigrams, bigrams, and trigrams during training. We also tried to improve the explicability of the model and to limit the black box effect [70,71] by displaying the outputs of the neural networks through an interactive word cloud interface. The word cloud representation is easily understandable and made it possible to consider the outputs attributed by the neural networks to each word according to sentiment intensity score. Our study has also several limitations. First, our method was developed on a data set of tweets in English and needs to be adapted for other languages [72] and assessed with other extensive data sets [49,73]. Another limitation is the finite set of inclusion keywords, resulting in a potential lack of information due to the total number of keywords used. Further works should concentrate on the diversification of keywords used to provide better sensibility. Furthermore, duplicate tweets and retweets were removed during preprocessing to limit the risk of overrepresenting one person's view, but this may have also led to underestimating the weights of some words. Second, class imbalance was checked before training, and early stopping was used to prevent the neural network from overfitting the data set. This resulted in good performance, with a model accuracy of 81%. Published studies have reported accuracies ranging from 48% to 91% [47,49,50] with the use of supervised learning techniques such as SVM, Naïve Bayes, logistic regression, or word2vec models. However, these performances were measured for binary sentiment classification (ie, negative vs positive sentiment). Here, we decided to consider neutral sentiments too, because it has been shown that tweets can be associated with neutral sentiments [74]. This choice allowed us to give more explicability and granularity but remains an issue because of our inability to compare our results with those of other studies.

Comparison With Prior Work

Use of Social Media to Capture Public Opinion

Approaches other than social media mining have been described. Focus groups provide a good understanding of public opinion and sentiments but are time-consuming and not necessarily representative of the whole population [4,6,75] as shown by Rowe et al [76] during the avian influenza crisis. Telephone and web-based surveys are expensive and time-consuming [77]. Systematic reviews analyze studies capturing public opinion [75] but are inappropriate in pandemic conditions as they require multiple skill sets (eg, experts on the topic, systematic review

methodologists) and are hardly usable for real-time monitoring. Unlike these approaches, social media mining captures a large range of opinions from a large sample, rapidly and for a reasonable cost [38,75]. It also has proven useful for understanding the attitudes and behavior of the public during a crisis [78]. For example, before the COVID pandemic, Chew et al [16] used Twitter to extract public perceptions of H1N1 during the H1N1 pandemic. However, some limitations are inherent to social media: The studied population is limited to social media users [79], the geographic location of users cannot be assumed with absolute certainty [80], and analyses are limited to a given language and source (eg, Twitter). Our study illustrates that, despite these issues, social media mining remains an efficient way to capture the thoughts, feelings, and fears of part of the population during a pandemic.

Research Perspectives

As the detection of topics and sentiments is directly related to neural network accuracy, more options could be explored to obtain higher scores, such as replacing word2vec embedding with Embeddings from Language Models (ELMo) [75] or Bidirectional Encoder Representation from Transformers (BERT) [14], which have proven useful for aspect-based sentiment classification [4,76]. The development of a Twitter-specific version of sentiment lexicons integrating web-specific elements such as emojis, abbreviations, or hashtags might also improve results [77]. Future research should concentrate on adding more granularity to the emotion expressed in tweets, by using emotion-specific lexicons to annotate the tweets with specific emotions such as fear, sadness, or happiness [21]. Newly developed initiatives such as the Linguistic Inquiry and Word Count (LIWC) dictionary [81] could also fulfill this task as they provide a dictionary able to recognize emotional words and automatically categorize them as more granular emotions in a hierarchical way (ie, each granular emotion, such as anger, is a child of a top-level emotion like a negative emotion).

Implications for Public Health

Our method could be used to guide public health decisions [77]. Besides factual parameters such as the disease characteristics or the burden it poses to the health care system [77], public opinion must also be considered to ensure that public health decisions are in line with the beliefs and priorities of the public [77]. Since many people use social media to share opinions and sentiments [79], they could provide policy makers and clinicians an opportunity to understand, in real time, the expectations, beliefs, and behaviors of the population and to adapt public health decisions accordingly [82,83]. They can also be used to communicate timely messages to the population [84] and thus to increase the chance of successful adoption of measures by the population. The development of indicators based on the real-time tracking of health-related conversations on social media is becoming crucial [9,85-87]. A major contribution of this study is to show the usefulness of deep learning methods to simultaneously capture public opinion and associated sentiments from large amounts of social media data.

Conclusions

We developed a new approach to conduct both sentiment and topic analyses on social media data by leveraging deep neural networks in conjunction with lexicons. We visualized the outputs of the neural network through a word cloud web interface displaying the weighted words associated with each sentiment intensity score. We demonstrated the utility of our method by applying it to a COVID-19 data set and identifying the main positive and negative topics discussed on Twitter in reaction to the announcement of the pandemic by the WHO. Future studies should concentrate on improving neural network performance and adding granularity to emotion detection. Our method may eventually prove useful for developing indicators for monitoring public opinion during pandemics.

Acknowledgments

We thank Delphine Colliot and Audrey Vignon for help with data collection. We thank the reviewers for help with manuscript improvement.

Data Availability

The data that support the findings of this study are not openly available due to Twitter identifying information and are available from the corresponding author upon reasonable request.

Authors' Contributions

A Boukobza and RT designed the study and the methodology, implemented the study, created the visualizations, analyzed the data, and wrote the original draft of the manuscript. A Boukobza, BR, and RT collected the data. All authors performed the validation and reviewed, edited, and approval the final manuscript.

Conflicts of Interest

None declared.

References

1. Madhav N, Oppenheim B, Gallivan M, Mulembakani P, Rubin E, Wolfe N. Pandemics: Risks, Impacts, and Mitigation. In: Jamison DT, Gelband H, Horton S, Jha P, Laxminarayan R, Mock CN, et al, editors. Disease Control Priorities: Improving Health and Reducing Poverty. Washington DC: The International Bank for Reconstruction and Development and The World Bank; 2017.
2. Novel coronavirus (2019-nCoV) situation report. World Health Organization. 2020 Jan 21. URL: https://apps.who.int/iris/handle/10665/330760 [accessed 2022-05-15]
3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020 Feb 20;382(8):727-733 [FREE Full text] [doi: 10.1056/NEJMoa2001017] [Medline: 31978945]
4. Teasdale E, Yardley L. Understanding responses to government health recommendations: public perceptions of government advice for managing the H1N1 (swine flu) influenza pandemic. Patient Educ Couns 2011 Dec;85(3):413-418. [doi: 10.1016/j.pec.2010.12.026] [Medline: 21295434]
5. Henrich N, Holmes B. The public's acceptance of novel vaccines during a pandemic: a focus group study and its application to influenza H1N1. Emerg Health Threats J 2009;2:e8 [FREE Full text] [doi: 10.3134/ehtj.09.008] [Medline: 22460289]
6. Morrison LG, Yardley L. What infection control measures will people carry out to reduce transmission of pandemic influenza? A focus group study. BMC Public Health 2009 Jul 23;9:258 [FREE Full text] [doi: 10.1186/1471-2458-9-258] [Medline: 19627568]
7. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 2020 Apr 21;22(4):e19016 [FREE Full text] [doi: 10.2196/19016] [Medline: 32287039]
8. Zhao Y, Cheng S, Yu X, Xu H. Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. J Med Internet Res 2020 May 04;22(5):e18825 [FREE Full text] [doi: 10.2196/18825] [Medline: 32314976]
9. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang Y. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc 2020 Aug 01;27(8):1310-1315 [FREE Full text] [doi: 10.1093/jamia/ocaa116] [Medline: 32620975]
10. Wang X, Zou C, Xie Z, Li D. Public opinions towards COVID-19 in California and New York on Twitter. medRxiv 2020 Jul 14:1 [FREE Full text] [doi: 10.1101/2020.07.12.20151936] [Medline: 32699856]
11. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach. J Med Internet Res 2020 Nov 25;22(11):e20550 [FREE Full text] [doi: 10.2196/20550] [Medline: 33119535]
12. Zou C, Wang X, Xie Z, Li D. Public reactions towards the COVID-19 pandemic on Twitter in the United Kingdom and the United States. medRxiv 2020 Jul 28:1 [FREE Full text] [doi: 10.1101/2020.07.25.20162024] [Medline: 32766599]
13. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social network analysis of COVID-19 sentiments: application of artificial intelligence. J Med Internet Res 2020 Aug 18;22(8):e22590 [FREE Full text] [doi: 10.2196/22590] [Medline: 32750001]
14. Chandrasekaran R, Mehta V, Valkunde T, Moustakas E. Topics, trends, and sentiments of Tweets about the COVID-19 pandemic: temporal infoveillance study. J Med Internet Res 2020 Oct 23;22(10):e22624 [FREE Full text] [doi: 10.2196/22624] [Medline: 33006937]
15. Fung IC, Duke CH, Finch KC, Snook KR, Tseng P, Hernandez AC, et al. Ebola virus disease and social media: A systematic review. Am J Infect Control 2016 Dec 01;44(12):1660-1671. [doi: 10.1016/j.ajic.2016.05.011] [Medline: 27425009]
16. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]
17. Akter S, Aziz MT. Sentiment analysis on facebook group using lexicon based approach. 2017 Presented at: 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT); September 22-24, 2016; Dhaka, Bangladesh. [doi: 10.1109/CEEICT.2016.7873080]
18. Ghiassi M, Skinner J, Zimbra D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications 2013 Nov;40(16):6266-6282. [doi: 10.1016/j.eswa.2013.05.057]
19. Ji X, Chun SA, Geller J. Monitoring Public Health Concerns Using Twitter Sentiment Classifications. 2013 Presented at: IEEE International Conference on Healthcare Informatics; September 9-11, 2013; Philadelphia, PA. [doi: 10.1109/ICHI.2013.47]
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for language understanding. arXiv. Preprint posted online on May 24, 2019 [FREE Full text] [doi: 10.18653/v1/n19-1423]
21. Geirhos R, Janssen D, Schütt H, Rauber J, Bethge M, Wichmann F. Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv. Preprint posted online on December 11, 2018 [FREE Full text] [doi: 10.48550/arXiv.1706.06969]
22. Wahbeh A, Nasralah T, Al-Ramahi M, El-Gayar O. Mining physicians' opinions on social media to obtain insights into COVID-19: mixed methods analysis. JMIR Public Health Surveill 2020 Jun 18;6(2):e19276 [FREE Full text] [doi: 10.2196/19276] [Medline: 32421686]

23. Park HW, Park S, Chong M. Conversations and medical news frames on Twitter: infodemiological study on COVID-19 in South Korea. J Med Internet Res 2020 May 05;22(5):e18897 [FREE Full text] [doi: 10.2196/18897] [Medline: 32325426]
24. Wang T, Huang Z, Gan C. On mining latent topics from healthcare chat logs. J Biomed Inform 2016 Jun;61:247-259 [FREE Full text] [doi: 10.1016/j.jbi.2016.04.008] [Medline: 27132766]
25. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach. J Med Internet Res 2020 Apr 28;22(4):e19118 [FREE Full text] [doi: 10.2196/19118] [Medline: 32302966]
26. Jo W, Lee J, Park J, Kim Y. Online information exchange and anxiety spread in the early stage of the novel coronavirus (COVID-19) outbreak in South Korea: structural topic model and network analysis. J Med Internet Res 2020 Jun 02;22(6):e19455 [FREE Full text] [doi: 10.2196/19455] [Medline: 32463367]
27. Ahne A, Orchard F, Tannier X, Perchoux C, Balkau B, Pagoto S, et al. Insulin pricing and other major diabetes-related concerns in the USA: a study of 46 407 tweets between 2017 and 2019. BMJ Open Diabetes Res Care 2020 Jun;8(1) [FREE Full text] [doi: 10.1136/bmjdrc-2020-001190] [Medline: 32503810]
28. Khan MT, Durrani M, Ali A, Inayat I, Khalid S, Khan KH. Sentiment analysis and the complex natural language. Complex Adapt Syst Model 2016 Feb 03;4(1). [doi: 10.1186/s40294-016-0016-9]
29. Drus Z, Khalid H. Sentiment analysis in social media and its application: systematic literature review. Procedia Computer Science 2019;161:707-714. [doi: 10.1016/j.procs.2019.11.174]
30. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
31. Hays R, Daker-White G. The care.data consensus? A qualitative analysis of opinions expressed on Twitter. BMC Public Health 2015 Sep 02;15:838 [FREE Full text] [doi: 10.1186/s12889-015-2180-9] [Medline: 26329489]
32. Huston P, Rowan M. Qualitative studies. Their role in medical research. Can Fam Physician 1998 Nov;44:2453-2458 [FREE Full text] [Medline: 9839063]
33. Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, et al. Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. PLoS One 2011 Mar 17;6(3):e18029 [FREE Full text] [doi: 10.1371/journal.pone.0018029] [Medline: 21437291]
34. Karami A, Gangopadhyay A, Zhou B, Kharrazi H. Fuzzy approach topic discovery in health and medical corpora. Int. J. Fuzzy Syst 2017 May 17;20(4):1334-1345. [doi: 10.1007/s40815-017-0327-9]
35. Armouty B, Tedmori S. Automated Keyword Extraction using Support Vector Machine from Arabic News Documents. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) Internet Amman, Jordan: IEEE; 2019 Presented at: IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT); April 9-11, 2019; Amman, Jordan. [doi: 10.1109/jeeit.2019.8717420]
36. Dumais ST. Latent semantic analysis. Ann. Rev. Info. Sci. Tech 2005 Sep 22;38(1):188-230. [doi: 10.1002/aris.1440380105]
37. Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R. Using latent semantic analysis to improve access to textual information. 1988 Presented at: CHI88: Human Factors in Computing Systems; May 15-19, 1988; Washington, DC. [doi: 10.1145/57167.57214]
38. Zhang H, Wheldon C, Dunn AG, Tao C, Huo J, Zhang R, et al. Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. J Am Med Inform Assoc 2020 Feb 01;27(2):225-235 [FREE Full text] [doi: 10.1093/jamia/ocz191] [Medline: 31711186]
39. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003;3:993-1022 [FREE Full text] [doi: 10.5555/944919.944937]
40. Wang S, Ding Y, Zhao W, Huang Y, Perkins R, Zou W, et al. Text mining for identifying topics in the literatures about adolescent substance use and depression. BMC Public Health 2016 Mar 19;16(1):279 [FREE Full text] [doi: 10.1186/s12889-016-2932-1] [Medline: 26993983]
41. Schmidt T, Burghardt M. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature 2018:139-149 [FREE Full text] [doi: 10.18653/v1/w17-22]
42. Mohammad S, Salameh M, Kiritchenko S. Sentiment Lexicons for Arabic Social Media. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) 2016:33-37 [FREE Full text]
43. Hu M, Liu B. Mining and Summarizing Customer Reviews. 2004 Presented at: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 22-25, 2004; Seattle, WA. [doi: 10.1145/1014052.1014073]
44. Nielsen F. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv. Preprint posted online March 15, 2011 [FREE Full text] [doi: 10.48550/arXiv.1103.2903]
45. Mohammad S, Turney P. Crowdsourcing a Word-Emotion Association Lexicon. arXiv. Preprint posted online on August 28, 2013 [FREE Full text] [doi: 10.1111/j.1467-8640.2012.00460.x]
46. Mohammad SM, Turney PD. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. 2010 Presented at: Workshop on Computational Approaches to Analysis and Generation of Emotion in Text; June 5, 2010; Los Angeles, CA. [doi: 10.4324/9780429508059-1]

47. Hassan AUI, Hussain J, Hussain M, Sadiq M, Lee S. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. 2017 Presented at: International Conference on Information and Communication Technology Convergence (ICTC); December 14, 2017; Jeju, South Korea. [doi: 10.1109/ictc.2017.8190959]
48. Chekima K, Alfred R. Sentiment Analysis of Malay Social Media Text. In: Alfred R, Iida H, Ag Ibrahim A, Lim Y, editors. Computational Science and Technology. Singapore: Springer; 2018:205-219.
49. Mamidi R, Miller M, Banerjee T, Romine W, Sheth A. Identifying key topics bearing negative sentiment on Twitter: insights concerning the 2015-2016 Zika epidemic. JMIR Public Health Surveill 2019 Jun 04;5(2):e11036 [FREE Full text] [doi: 10.2196/11036] [Medline: 31165711]
50. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'Bad' is 'Good'": identifying personal communication and sentiment in drug-related tweets. JMIR Public Health Surveill 2016 Oct 24;2(2):e162 [FREE Full text] [doi: 10.2196/publichealth.6327] [Medline: 27777215]
51. Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. HP Laboratories. 2011. URL: https://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf [accessed 2022-05-15]
52. Lin C, He Y. Joint sentiment/topic model for sentiment analysis. 2009 Presented at: CIKM '09: Conference on Information and Knowledge Management; November 2-6, 2009; Hong Kong. [doi: 10.1145/1645953.1646003]
53. Lin C, He Y, Everson R, Ruger S. Weakly supervised joint sentiment-topic detection from text. IEEE Trans. Knowl. Data Eng 2012 Jun;24(6):1134-1145. [doi: 10.1109/tkde.2011.48]
54. Jo Y, Oh AH. Aspect and sentiment unification model for online review analysis. 2011 Presented at: WSDM'11: Fourth ACM International Conference on Web Search and Data Mining; February 9-12, 2011; Hong Kong. [doi: 10.1145/1935826.1935932]
55. Mei Q, Ling X, Wondra M, Su H, Zhai CX. Topic sentiment mixture: modeling facets and opinions in weblogs. 2007 Presented at: Topic sentiment mixture: modeling facets and opinions in weblogs; May 8-12, 2007; Banff, Alberta, Canada p. A. [doi: 10.1145/1242572.1242596]
56. Dermouche M, Kouas L, Velcin J, Loudcher S. A joint model for topic-sentiment modeling from text. 2015 Presented at: SAC 2015: Symposium on Applied Computing; April 13-17, 2015; Salamanca, Spain. [doi: 10.1145/2695664.2695726]
57. Dermouche M, Velcin J, Khouas L, Loudcher S. A Joint Model for Topic-Sentiment Evolution over Time. 2014 Presented at: IEEE International Conference on Data Mining; December 14-17, 2014; Shenzhen, China. [doi: 10.1109/icdm.2014.82]
58. Giménez M, Palanca J, Botti V. Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis. Neurocomputing 2020 Feb;378:315-323. [doi: 10.1016/j.neucom.2019.08.096]
59. Park H, Song M, Shin K. Deep learning models and datasets for aspect term sentiment classification: Implementing holistic recurrent attention on target-dependent memories. Knowledge-Based Systems 2020 Jan;187:104825. [doi: 10.1016/j.knosys.2019.06.033]
60. Abd Elaziz M, Al-qaness MAA, Ewees AA, Dahou A. Recent Advances in NLP: The Case of Arabic Language. In: Kacprzyk J, editor. Studies in Computational Intelligence. Cham, Switzerland: Springer International Publishing; 2020.
61. twintproject / twint. GitHub. 2021 Mar 02. URL: https://github.com/twintproject/twint [accessed 2022-05-15]
62. Porter MF. An algorithm for suffix stripping. Program: electronic library & information systems 2006;40(3):211-218. [doi: 10.1108/00330330610681286]
63. Natural Language Toolkit. URL: http://nltk.org/ [accessed 2022-05-15]
64. De Queiroz G, Fay C, Hvitfeldt E, Keyes O, Misra K, Mastny T, et al. tidytext: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools. CRAN.R. 2022 May 09. URL: https://CRAN.R-project.org/package=tidytext [accessed 2022-05-15]
65. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014:1532-1543. [doi: 10.3115/v1/d14-1162]
66. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. IEEE 1998;86(11):2278-2324. [doi: 10.1109/5.726791]
67. Chang W, Cheng J, Allaire J, Xie Y, McPherson J, RStudio. shiny: Web Application Framework for R. CRAN.R. 2021 Oct 02. URL: https://CRAN.R-project.org/package=shiny [accessed 2022-05-15]
68. Tweeter Word Cloud Generator. URL: https://adrien-boukobza.shinyapps.io/word_cloud_shiny/ [accessed 2022-05-15]
69. Rinker T. textstem: Tools for Stemming and Lemmatizing Text. CRAN.R. 2018 Apr 09. URL: https://CRAN.R-project.org/package=textstem [accessed 2022-05-15]
70. Castelvecchi D. Can we open the black box of AI? Nature. 2016 Oct 5. URL: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731 [accessed 2022-05-15]
71. Lamy J, Sedki K, Tsopra R. Explainable decision support through the learning and visualization of preferences from a formal ontology of antibiotic treatments. J Biomed Inform 2020 Apr;104:103407 [FREE Full text] [doi: 10.1016/j.jbi.2020.103407] [Medline: 32156641]
72. AL-Rubaiee H, Qiu R, Li D. The importance of neutral class in sentiment analysis of Arabic tweets. IJCSIT 2016 Apr 30;8(2):17-31. [doi: 10.5121/ijcsit.2016.8202]

73. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018 Mar;286(3):800-809. [doi: 10.1148/radiol.2017171920] [Medline: 29309734]
74. Kent EE, Prestin A, Gaysynsky A, Galica K, Rinker R, Graff K, et al. "Obesity is the New Major Cause of Cancer": connections between obesity and cancer on Facebook and Twitter. J Cancer Educ 2016 Sep 14;31(3):453-459. [doi: 10.1007/s13187-015-0824-1] [Medline: 25865399]
75. Giles EL, Adams JM. Capturing public opinion on public health topics: a comparison of experiences from a systematic review, focus group study, and analysis of online, user-generated content. Front Public Health 2015 Aug 24;3:200 [FREE Full text] [doi: 10.3389/fpubh.2015.00200] [Medline: 26380248]
76. Rowe G, Hawkes G, Houghton J. Initial UK public reaction to avian influenza: Analysis of opinions posted on the BBC website. Health, Risk & Society 2008 Aug;10(4):361-384. [doi: 10.1080/13698570802166456]
77. Lipsitch M, Finelli L, Heffernan RT, Leung GM, Redd SC, 2009 H1n1 Surveillance Group. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza A/H1N1. Biosecur Bioterror 2011 Jun;9(2):89-115 [FREE Full text] [doi: 10.1089/bsp.2011.0007] [Medline: 21612363]
78. Barros JM, Duggan J, Rebholz-Schuhmann D. The application of internet-based sources for public health surveillance (infoveillance): systematic review. J Med Internet Res 2020 Mar 13;22(3):e13680 [FREE Full text] [doi: 10.2196/13680] [Medline: 32167477]
79. Chou WS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: implications for health communication. J Med Internet Res 2009 Nov 27;11(4):e48 [FREE Full text] [doi: 10.2196/jmir.1249] [Medline: 19945947]
80. Burton SH, Tanner KW, Giraud-Carrier CG, West JH, Barnes MD. "Right time, right place" health communication on Twitter: value and accuracy of location information. J Med Internet Res 2012 Nov 15;14(6):e156 [FREE Full text] [doi: 10.2196/jmir.2121] [Medline: 23154246]
81. LIWC. URL: https://www.liwc.app/ [accessed 2022-05-15]
82. Dandala B, Joopudi V, Devarakonda M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. Drug Saf 2019 Jan 16;42(1):135-146. [doi: 10.1007/s40264-018-0764-x] [Medline: 30649738]
83. Vydiswaran VGV, Romero DM, Zhao X, Yu D, Gomez-Lopez I, Lu JX, et al. Uncovering the relationship between food-related discussion on Twitter and neighborhood characteristics. J Am Med Inform Assoc 2020 Feb 01;27(2):254-264 [FREE Full text] [doi: 10.1093/jamia/ocz181] [Medline: 31633756]
84. Liao Q, Yuan J, Dong M, Yang L, Fielding R, Lam WWT. Public engagement and government responsiveness in the communications about COVID-19 during the early epidemic stage in China: infodemiology study on social media data. J Med Internet Res 2020 May 26;22(5):e18796 [FREE Full text] [doi: 10.2196/18796] [Medline: 32412414]
85. Yeung D. Social media as a catalyst for policy action and social change for health and well-being: viewpoint. J Med Internet Res 2018 Mar 19;20(3):e94 [FREE Full text] [doi: 10.2196/jmir.8508] [Medline: 29555624]
86. McClellan C, Ali MM, Mutter R, Kroutil L, Landwehr J. Using social media to monitor mental health discussions - evidence from Twitter. J Am Med Inform Assoc 2017 May 01;24(3):496-502 [FREE Full text] [doi: 10.1093/jamia/ocw133] [Medline: 27707822]
87. Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. J Am Med Inform Assoc 2019 Dec 01;26(12):1618-1626 [FREE Full text] [doi: 10.1093/jamia/ocz156] [Medline: 31562510]

Abbreviations

ASUM: Aspect and Sentiment Unification Model
BERT: Bidirectional Encoder Representation from Transformers
CNN: convolutional neural network
ELMo: Embeddings from Language Models
GloVe: Global Vector for word representation
JST: joint sentiment topic
LDA: latent Dirichlet allocation
LIWC: Linguistic Inquiry and Word Count
NLP: natural language processing
NLTK: Natural Language Toolkit
SVM: support vector machine
TSM: Topic-Sentiment Mixture
TTTS: Time-aware Topic Sentiment
WHO: World Health Organization

Edited by C Lovis; submitted 16.10.21; peer-reviewed by J Chen, R Benson; comments to author 31.01.22; revised version received 14.02.22; accepted 21.04.22; published 25.05.22.

Please cite as:
Boukobza A, Burgun A, Roudier B, Tsopra R
Deep Neural Networks for Simultaneously Capturing Public Topics and Sentiments During a Pandemic: Application on a COVID-19 Tweet Data Set
JMIR Med Inform 2022;10(5):e34306
URL: https://medinform.jmir.org/2022/5/e34306
doi:10.2196/34306
PMID:35533390

Original Paper

Domain-Specific Common Data Elements for Rare Disease Registration: Conceptual Approach of a European Joint Initiative Toward Semantic Interoperability in Rare Disease Research

Haitham Abaza[1], PhD; Dennis Kadioglu[1], MSc; Simona Martin[2], PhD; Andri Papadopoulou[2], PhD; Bruna dos Santos Vieira[3,4], MSc; Franz Schaefer[5], Dr med, Prof Dr; Holger Storf[1], Prof Dr

[1]Institute of Medical Informatics, Goethe University Frankfurt, University Hospital Frankfurt, Frankfurt am Main, Germany

[2]European Commission, Joint Research Centre, Ispra, Italy

[3]Department of Medical Imaging, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, Netherlands

[4]Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, Netherlands

[5]Division of Pediatric Nephrology, Center for Pediatrics and Adolescent Medicine, Heidelberg, Germany

Corresponding Author:
Dennis Kadioglu, MSc
Institute of Medical Informatics
Goethe University Frankfurt
University Hospital Frankfurt
Theodor-Stern-Kai 7
Frankfurt am Main, 60590
Germany
Phone: 49 696301 ext 80692
Email: dennis.kadioglu@kgu.de

Abstract

Background: With hundreds of registries across Europe, rare diseases (RDs) suffer from fragmented knowledge, expertise, and research. A joint initiative of the European Commission Joint Research Center and its European Platform on Rare Disease Registration (EU RD Platform), the European Reference Networks (ERNs), and the European Joint Programme on Rare Diseases (EJP RD) was launched in 2020. The purpose was to extend the set of common data elements (CDEs) for RD registration by defining domain-specific CDEs (DCDEs).

Objective: This study aims to introduce and assess the feasibility of the concept of a joint initiative that unites the efforts of the European Platform on Rare Disease Registration Platform, ERNs, and European Joint Programme on Rare Diseases toward extending RD CDEs, aiming to improve the semantic interoperability of RD registries and enhance the quality of RD research.

Methods: A joint conference was conducted in December 2020. All 24 ERNs were invited. Before the conference, a survey was communicated to all ERNs, proposing 18 medical domains and requesting them to identify highly relevant choices. After the conference, a 3-phase plan for defining and modeling DCDEs was drafted. Expected outcomes included harmonized lists of DCDEs.

Results: All ERNs attended the conference. The survey results indicated that genetic, congenital, pediatric, and cancer were the most overlapping domains. Accordingly, the proposed list was reorganized into 10 domain groups and recommunicated to all ERNs, aiming at a smaller number of domains.

Conclusions: The approach described for defining DCDEs appears to be feasible. However, it remains dynamic and should be repeated regularly based on arising research needs.

(JMIR Med Inform 2022;10(5):e32158) doi:10.2196/32158

KEYWORDS

semantic interoperability; common data elements; standardization; data collection; data discoverability; rare diseases; EJP RD; EU RD Platform; ERNs; FAIRification; health infrastructure; industry; medical informatics; health platforms; health registries; health and research platforms; health domains

Introduction

Background

Patient registries and databases are fundamental instruments for increasing knowledge on rare diseases (RDs), supporting clinical, epidemiological, and basic research, and improving patient care and health care planning [1,2]. With >600 registries across Europe [3], RDs suffer from fragmented knowledge, scattered expertise, and research duplication [1]. Data are not collected in a uniform way throughout Europe, and there are no shared standards to analyze the information [3]. The use of international coding and nomenclature, minimal common data sets, and good practice guidelines enhances the interoperability and maximizes the utility of RD registries. This allows data to be efficiently pooled to reach sufficient sample sizes for clinical and public health research focusing on disease etiology, pathogenesis, diagnosis, and therapy [1,2].

On the Rare Disease Day (2019), the European Commission announced a new web-based knowledge-sharing platform to promote better diagnosis and treatment for >30 million patients with RD. Developed by the Joint Research Centre (JRC) of the European Commission, the EU RD Platform aims to bring together European RD registries, thus overcoming fragmentation and promoting interoperability between existing and new registries. Moreover, the platform seeks to standardize data collection and data exchange at the EU level, thereby supporting quality RD research, enhancing diagnosis and treatment outcomes, and improving the lives of patients and their families [4]. The efforts of European Reference Networks (ERNs) toward establishing ERN-wide registries are also implicitly supported by the platform [3], mainly by offering a patient pseudonymization and privacy-preserving linkage service.

By delivering EU standards for data collection and data sharing, the EU RD Platform is a significant asset for the European Joint Programme on Rare Diseases (EJP RD) [5], which aims to establish an innovation network for rapidly translating research results into clinical and health care applications [3]. The EJP RD brings together over 130 institutions from 35 countries to collaboratively build infrastructure and digital platforms, which promote cross-border sharing of clinical data and expertise. The ultimate goal is to overcome the fragmentation of RD resources and to foster RD care and medical innovation. The EJP RD also aims to use, support, and connect already-funded tools operating within the field of RD research and adapt them to the needs of end users through implementation tests in real settings [6]. Through EJP RD, the EU RD Platform resources can be disseminated to future research projects and exposed to a wider community of RD researchers, clinicians, and patients in Europe and elsewhere [3].

Aiming to make RD registries and their data searchable and findable, the EU RD Platform comprises the European Rare Disease Registry Infrastructure (ERDRI) [7], which includes the European Directory of Registries (ERDRI.dor), the Central Metadata Repository (ERDRI.mdr), and the pseudonymization tool. Details on their infrastructure and functioning will be published elsewhere. However, we focus here on the set of common data elements (CDEs) for RD registration [8], which is another important building block of the platform. Developed by experts from various EU projects (eg, European Union Committee of Experts on Rare Diseases Joint Action [EUCERD], European Platform for Rare Disease Registries [EPIRARE], and RD-Connect) related to common data sets, the set of CDEs was released by the EU RD Platform as the first practical instrument toward increasing the interoperability of RD registries [9]. The set recommends the collection of 16 data elements by all European RD registries, as they are considered essential for RD research. The 16 CDEs are classified into various groups, including personal data, diagnosis, disease history, care pathway, information for research purposes, and a disability profile. Exemplary CDEs include age (date of birth), sex (male, female, undetermined, or fetus), status (alive, dead, lost to follow-up, or opted-out), and RD diagnosis (ORPHAcode) [9].

Although CDEs constitute a common basis for characterizing patients with RD across all 24 ERNs, many overlaps between ERN domains are not clearly defined. For instance, there are 3 oncological (ERN PaedCan, ERN EURACAN, and ERN EuroBloodNet) and 3 neurological (ERN EpiCare, ERN-RND, and ERN EURO-NMD) ERNs, among others, with numerous diseases covered by each of them being treated jointly. The list of the 24 initially funded ERNs that have been considered in the context of this work can be found in Multimedia Appendix 1. Furthermore, beyond CDEs, many ERN registries collect data elements that may be commonly used by others working in the same domain. However, no standards exist for categorizing such commonalities. Domain-specific CDEs (DCDEs) are designed for use in studies or registries of a particular topic, disease or condition, body system, or other classifications (eg, cancer, Parkinson disease, Alzheimer disease, diabetes, or ophthalmology). Some domains are broadly applicable to a wide range of studies, whereas others are more useful in specific fields of clinical research [10]. Therefore, the definition of DCDEs for the various RD domains is expected to standardize data collection, thus enhancing the interoperability and facilitating the discoverability of data stored in RD registries.

In 2019, the EJP RD formed an expert workforce to assist the ERN Registry Task Force (TF) on interoperability and standardization issues [11,12]. Extracted from the data dictionaries of the first 4 ERN registries (ERKReg [ERKNet], U-IMD [MetabERN], EURRECA [ENDO-ERN], and DATA WAREHOUSE [ERN-LUNG]), a Common Data Dictionary (CDD) was introduced as a tool to avoid fragmentation and ensure registry interoperability. Accordingly, the TF committed to the use of the CDD as part of the group's efforts to develop ERN registries in full compliance with the FAIR principles (findability, accessibility, interoperability, and reusability). Therefore, these efforts are expected to improve research transparency and facilitate knowledge discovery for both humans and machines [11,13].

Rationale

To achieve semantic interoperability between RD registries, a joint initiative of the EU RD Platform, the ERN Registry TF, and the EJP RD registry interoperability work focus group was

launched in 2020. Driven by the research needs of ERNs, the purpose was to extend the set of CDEs for RD registration by defining DCDEs, that is, ones that are considered necessary within each particular ERN domain. This idea was first expressed at an ERN Registry TF meeting in Brussels around mid-2019, when some ERNs indicated that they had already collected a small number of data elements that commonly exist in registries of their domain.

To this end, a joint conference took place in December 2020, bringing together the EU RD Platform team members, ERN representatives and registry owners, and EJP RD partners to discuss the concept of DCDEs cooperatively. The conference also aimed to tackle core questions, such as whether all ERNs already had a defined data set in place, and for which reasons they believed DCDEs would be necessary. Medical experts were also intended to indicate, during and after the conference, whether some domains could already be derived from existing overlaps and consider experts who could take charge of such domains. Exploring how harmonized lists of DCDEs could be produced was planned at a later stage, as well as identifying appropriate standards, ontologies, and terminologies to finally annotate DCDEs and integrate them into CDE semantic modeling activities of EJP RD.

Led by the ERN Registry TF, initial efforts were already made by 4 ERNs (ERKNet, MetabERN, ENDO-ERN, and ERN-LUNG) toward the creation of a CDD, mainly collecting common data fields among their registries in an Excel (Microsoft Inc) table. In continuation to these efforts, the plan is to have medical experts from all 24 ERNs drive this initiative toward the following:

- Forming ERN domain groups
- Defining DCDEs for each domain group by identifying commonalities among relevant ERNs
- Harmonizing, modeling, and publishing DCDEs and adding them to ERDRI.mdr

Objectives

This study introduces the concept of a collaborative initiative, which aims to prevent duplicated efforts by uniting and coordinating the activities of the EU RD Platform, ERN Registry TF, and EJP RD on topics such as the CDEs, CDD, and common metadata and data model of the EJP RD. Moreover, the initiative aims to further standardize RD registration by extending the set of CDEs with DCDEs, thereby improving the semantic interoperability of RD registries and enhancing the quality of RD research. The study also assesses the feasibility of the concept by examining previous efforts to define (D)CDEs and exploring if and how DCDEs can benefit the RD field from the perspective of the ERNs.

Methods

Conference Participants

All 24 ERNs were invited to attend the conference. Speakers included participants from the JRC/EU RD Platform as well as EJP RD experts from various backgrounds, particularly focusing on common data sets and FAIRification. ERN representatives were preferably required to have a medical background and considerable involvement in registry activities. These were considered necessary requirements to identify essential ERN domains as well as existing overlaps, if any, thus paving the way for creating lists of DCDEs. The EJP RD previously built a database of experts involved in registry design and construction, listing their names, institutions, contact details, and expertise. Although not yet complete, the database included experts from all 3 parties (JRC/EU RD Platform, ERNs, and EJP RD) who indicated working with registries. Therefore, it was intended for use, together with other resources, to identify appropriate participants for the next steps.

Preconference Tasks

Before the conference, a Forms (Microsoft Inc) survey was prepared and communicated to all 24 ERNs through the FAIRification stewards of the EJP RD. The survey proposed a list of medical domains and requested ERNs to identify those that generally fit their activities. The list comprised 18 domains, mainly suggesting the specialties indicated in the name of each ERN (eg, ERN-EYE—Sight, ERKNet—Renal, and EURACAN—Cancer). The main survey item read, “To which domain(s) do you think your ERN fits?” and enabled checking multiple answers. Another optional item asked if any of the suggested domains could be grouped together and allowed for text answers (eg, cancer and congenital). The deadline for completing the survey was set on the day of the conference. However, we also planned to collect any missing answers during or shortly after the conference.

The Conference

Organized by the EJP RD, the conference comprised three 40-minute sessions, the second of which was dedicated to DCDEs. The first presentation was held by the EU RD Platform team, providing some findings collected in preparation for their originally planned ERN workshop in March 2020. Unfortunately, this event was cancelled because of the COVID-19 pandemic. The team also expressed interest in having 2 specific questions answered, namely, whether each of the ERNs already had their data set in place (at the ERN level) and for which specific purpose they believed DCDEs were necessary. The survey results were then presented by EJP RD experts, giving some exemplary purposes to illustrate the importance of DCDEs and accordingly suggesting a scoring method for rating the importance of every identified DCDE within a particular domain. It was also indicated that a technical phase would follow the definition of DCDEs. Therefore, both EJP RD and EU RD Platform experts would guide the ERNs through harmonizing and modeling identified DCDEs, in preparation for adding them to ERDRI.mdr and extending semantic data modeling activities of the EJP RD.

Postconference Tasks

Following the conference, the remaining ERNs, which had not completed the survey, were requested to provide their answers, and the following 3-phase plan was jointly drafted:

1. Formation of ERN domain groups
    - EJP RD experts group suggested domains and request ERNs to review the groups

- EJP RD experts request ERNs to reorganize the suggested groups, if necessary, aiming for a minimum number of domain groups and a minimum number of ERNs per group
- ERNs elect relevant experts/curation team members for each group

2. Definition of DCDEs
   - Elected ERN experts suggest, curate, and define DCDEs by comparing their data dictionaries and identifying relevant commonalities for every domain group (the EJP RD can support if the data dictionaries are shared)
   - The EU RD Platform and EJP RD experts offer support by providing necessary templates for DCDE lists, scheduling domain meetings, and ensuring that everything is harmonized
   - The EU RD Platform and EJP RD experts prioritize DCDEs using the scoring method proposed by EJP RD, in case long lists are identified
3. Technical phase
   - EU RD Platform and EJP RD experts guide ERNs to extend the semantic data modeling activities of the EJP RD: harmonization, modeling, and mapping of DCDEs
   - EU RD Platform and EJP RD experts guide the ERNs in publishing DCDEs alongside CDEs and inclusion in ERDRI.mdr
   - New registries implement both CDEs and DCDEs

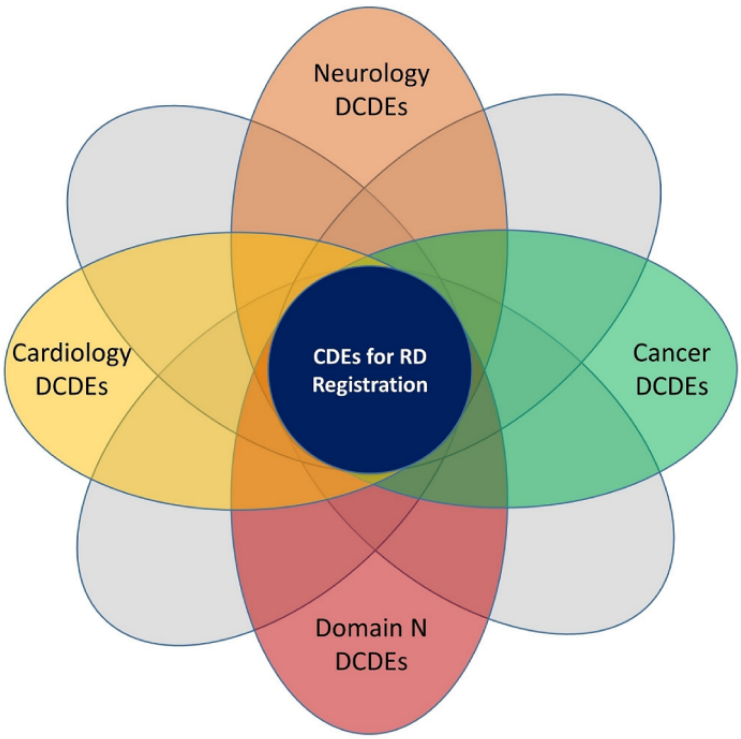
Scoring Method

Completed DCDE lists were planned to be sent to medical experts for review using a structured feedback method, allowing them to rate an arranged set of statements designed to indicate the relevance of each DCDE within a certain domain group. To ease the adoption of identified DCDEs by all RD registries, it was initially recommended that the rating statements address the following aspects: (1) importance of each DCDE for the integrity of a registry within a certain domain group, (2) reliability of data collection in each DCDE, (3) necessity of a DCDE for the analysis of the primary outcome of the registry, and (4) the time and cost required to collect each DCDE [14]. Other categories that might arise in discussions during or after the conference were also to be incorporated. On the basis of the feedback of experts, individual scores could eventually be calculated for every identified DCDE, thus reflecting the importance of each DCDE within each domain group. These scores could also be used by curation teams to prioritize their DCDEs, if their efforts culminated in prolonged lists.

Domain Representation

To visually represent the domains, several diagrams were prepared and circulated before and during the conference. Figure 1 illustrates this concept by showing domain overlaps and classifying DCDEs. Neurology, cancer, and cardiology were used for exemplary purposes, with *N* suggesting an unknown number of domains expected to be identified by the initiative. The CDEs of the EU RD Platform were placed in the center, ensuring that they would remain the basis for all DCDE lists.

Figure 1. Domain-specific common data elements (DCDEs) classification and domain overlaps. CDE: common data element; RD: rare disease.

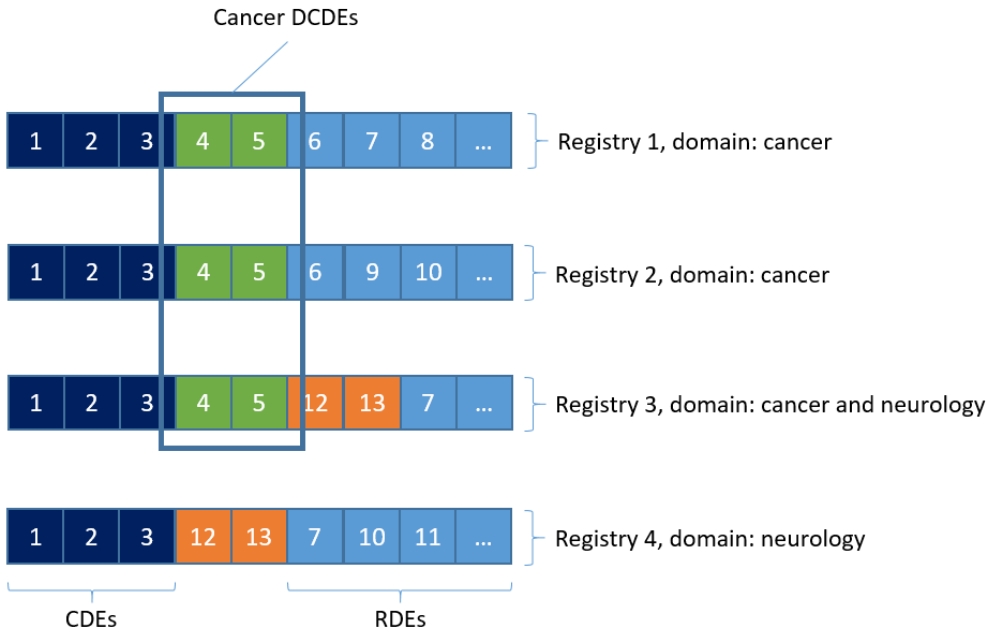


To illustrate the concept further, and how it is intended to classify registries, Figure 2 depicts an example of 4 ERN registries belonging to 2 exemplary domains, namely cancer and neurology. Numbered blocks were used to represent the data elements of a registry, whereas colors were applied to characterize the different types of data elements. In an ideal world, the registry data elements could be classified into the following:

1. CDEs, that is, ones that would commonly exist in all RD registries
2. DCDEs, that is, ones that would commonly exist in all registries of a particular RD domain
3. Registry-specific data elements, that is, ones considered specific for each registry's purpose, thus would commonly differ from one registry to another

This classification was also illustrated by the occurrence of navy blue data elements (CDEs) among all registries, regardless of their domain, as opposed to the green and orange elements present only in registries of the cancer and neurology domains, respectively. Furthermore, the presence of both the green and orange elements in registry 3 was used to indicate that a registry could belong to more than one domain, thereby incorporating DCDEs of both the cancer and neurology domains. Notably, the marked cancer DCDEs ensured that CDEs would constantly remain the basis for all domains while complemented by domain-specific extensions.

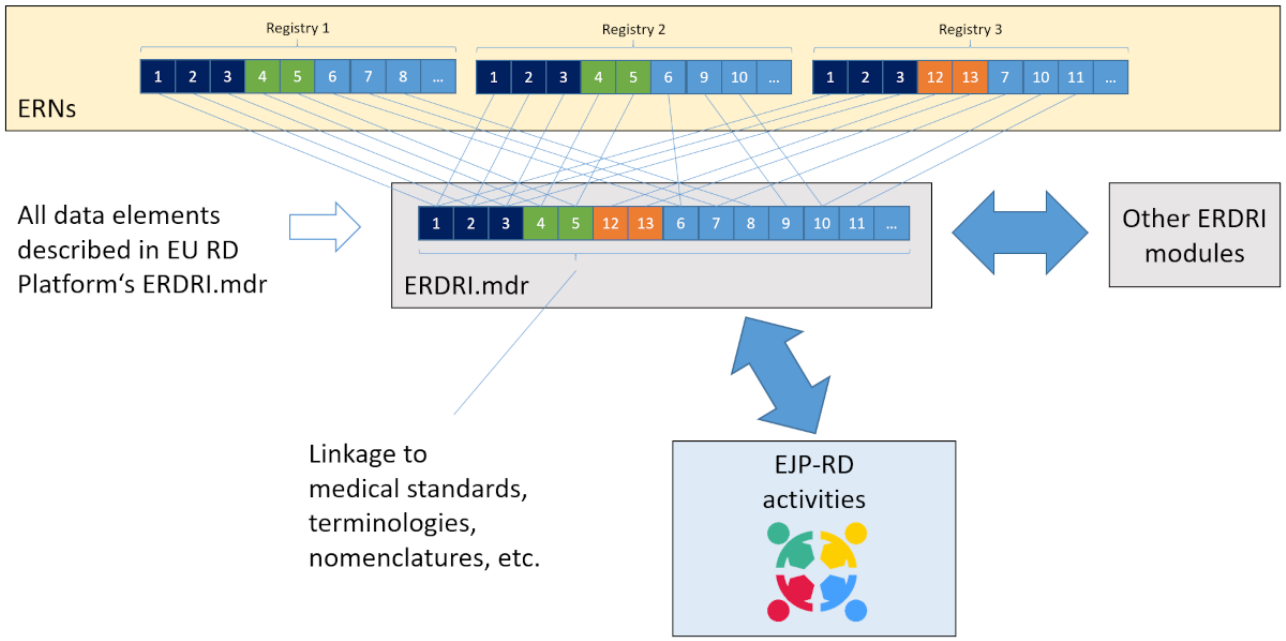
Figure 2. Classification of registries and their data elements. CDE: common data element; DCDE: domain-specific common data element; RDE: registry-specific data element.



Expected Outcomes

Lists of DCDEs for every identified ERN domain group were set as expected outcomes for the first and second phases. The technical phase was planned to harmonize these lists, thus removing duplicates, if any, and ensuring identical definitions of DCDEs and their values. The harmonized version would then be published in PDF format on the EU RD Platform and added to ERDRI.mdr to allow reusability. EJP RD experts would also explore existing ontologies and terminologies for annotating and incorporating the harmonized lists into the registry codebook of the EJP RD [15] and CDE semantic data model [16]. This is expected to enhance the interoperability of RD registries and boost RD research, which is the main objective of this project. Figure 3 shows a visualization of the concept, portraying the connections among the different members of the initiative.

Figure 3. Technical vision. CDE: common data element; EJP RD: European Joint Programme on Rare Diseases; ERDRI: European Rare Disease Registry Infrastructure; ERN: European Reference Network; EU RD Platform: European Platform on Rare Disease Registration; MDR: Metadata Repository.



Ethics Approval

Any data examined do not relate to specific individuals, which means that no personal harm can occur to individuals. Accordingly, the review by an ethics committee was not required.

Results

Results Overview

All 24 ERNs attended the conference. Most participants indicated that they already had their data sets in place. However, the main purpose for defining DCDEs could not be recognized. The EJP RD experts, however, presented three potential purposes that emphasized the importance of DCDEs: increasing interoperability, allowing data comparisons in joint research projects, and improving data discoverability.

A Scoring Method for the Nomination of Domain-Specific Common Data Elements

A scoring method was also presented, suggesting a way for medical experts to rate the importance of each DCDE based on these 3 purposes. Figure 4 shows the proposed scoring system using a 4-point scale. On the basis of the feedback of 3 experts, exemplary scores were also provided for the 2 DCDEs belonging to the cancer domain. For each item, the average of all expert scores was calculated to provide an overall item score. The overall DCDE score was then determined as the average of all 3 overall item scores. In this particular example, the score of DCDE 1 slightly exceeded that of DCDE 2, suggesting that it is somewhat more important for the cancer domain. Similarly, individual tables are meant to be constructed for each of the domains identified by the initiative.

Figure 4. Proposed scoring system and example. DCDE: domain-specific common data element; ERN: European Reference Network.

Item/Score	4	3	2	1
This data element is commonly collected by various ERNs within this domain and thus is necessary for improving interoperability.	Strongly agree	Agree	Fairly agree	Disagree
This data element captures sufficient detail for research purposes and thus is necessary for carrying out comparisons in joint research projects within this domain.				
This data element is commonly used to describe patients and thus is necessary for various purposes of discoverability within this domain.				

Cancer DCDEs	Item	Feedback			Overall item score	Overall DCDE score (4)
		Expert 1	Expert 2	Expert 3		
DCDE 1	This data element is commonly collected by various ERNs within this domain and thus is necessary for improving interoperability.	4	3	2	3	2.67
	This data element captures sufficient detail for research purposes and thus is necessary for carrying out comparisons in joint research projects within this domain.	3	2	4	3	
	This data element is commonly used to describe patients and thus is necessary for various purposes of discoverability within this domain.	1	2	3	2	
DCDE 2	This data element is commonly collected by various ERNs within this domain and thus is necessary for improving interoperability.	2	2	3	2.33	2.56
	This data element captures sufficient detail for research purposes and thus is necessary for carrying out comparisons in joint research projects within this domain.	4	3	1	2.67	
	This data element is commonly used to describe patients and thus is necessary for various purposes of discoverability within this domain.	3	1	4	2.67	

First Proposal of Domains and Domain Groups

EJP RD experts also presented the survey results, indicating the genetic, congenital, pediatric, and cancer domains as the most overlapping and adding to the importance of defining DCDEs. For the first survey item, which requested ERNs to select relevant domains from a list of 18 suggestions, responses from 22 (92%) ERNs were received and are presented in Table 1. The second survey item was answered by 14 (58%) ERNs. However, no patterns could be identified in the suggested domain groups.

Following the conference, the EJP RD experts and FAIRification stewards grouped some of the suggested domains, aiming at a minimum number of domain groups. The initially proposed domains were reorganized into 10 domain groups, and the survey answers were used to identify relevant ERNs. Table 2 shows the suggested list of domain groups as well as the corresponding ERNs. As shown, a single domain group could comprise multiple related ERNs and a single ERN could belong to various relevant domain groups. The list was communicated once more to all ERNs, requesting them to edit, merge, add, or move their ERN between domains as they saw necessary. They were also requested to elect, for every domain group, a person in charge and members of a data element curation team. This constitutes selected members, from every ERN of a particular domain group, intended to be in charge of drafting a DCDEs list, sending it to medical experts for review using the aforementioned scoring method, and accordingly agreeing on final definitions.

Table 1. Survey responses.

Domain	ERNs[a], n (%)
Genetic	20 (83)
Pediatrics	18 (75)
Congenital	14 (58)
Cancer	7 (29)
Endocrine and metabolism	6 (25)
Neurology	6 (25)
Renal	6 (25)
Immune disorders	6 (25)
Skin	5 (21)
Gastroenterology and hepatology	4 (17)
Lung	4 (17)
Other	4 (17)
Muscle, bone, and skeletal diseases	3 (13)
Cardiovascular	3 (13)
Hematological disorders	3 (13)
Urology	3 (13)
Respiratory	2 (8)
Head and neck	1 (4)
Sight	1 (4)

[a]ERN: European Reference Network.

Table 2. Suggested domain groups.

Domain group	Relevant ERNs[a]
Genetic	ERN-EYE, RARE-LIVER, ERNICA, ERKNet, ERN ITHACA, ERN GUARD-Heart, EPICARE, PaedCan, VASCERN, EuroBloodNet, Endo-ERN and ERN-BOND, MetabERN, ERN-TransplantChild, ERN-Skin, ERN CRANIO, ERN GENTURIS, RITA, and ERN Eurogen
Congenital	ERN-EYE, ERNICA, ERN ITHACA, VASCERN, Endo-ERN and ERN-BOND, MetabERN, ERN-TransplantChild, ERN-LUNG, ERN-Skin, ERN CRANIO, RITA, and ERN eUROGEN
Pediatrics	ERN-EYE, RARE-LIVER, ERNICA, ERN ITHACA, EPICARE, PaedCan, VASCERN, Endo-ERN and ERN-BOND, MetabERN, ERN-TransplantChild, ERN-LUNG, ERN-Skin, ERN CRANIO, ERN GENTURIS, RITA, ERN eUROGEN, and ERKNet
Cancer	RARE-LIVER, PaedCan ERN, ERN-EuroBloodNet, Endo-ERN and ERN-BOND, eUROGEN, ERN-Skin, and ERN GENTURIS
Neurological	ERKNet, EPICARE, ERN-RND, MetabERN, ITHACA, ERN Eurogen, and EURO-NMD
Immune and blood	RARE-LIVER, EPICARE, ERN-EuroBloodNet, ERN-TransplantChild, ERN-Skin, RITA, ERKNet, VASCERN, and ReConnet
Renal and urological	ERKNet, EPICARE, ERN-RND, Endo-ERN and ERN-BOND, MetabERN, ITHACA, ERN eUROGEN, ERN-EuroBloodNet, and ERN-TransplantChild
Respiratory and lung	VASCERN, ERN-EuroBloodNet, ERN-TransplantChild, ERN-LUNG, and ERNICA
Surgical	eUROGEN, ERN CRANIO, and ERN-TransplantChild
Muscle, bone, and skeletal	Endo-ERN and ERN-BOND, ITHACA, RITA, and EURO-NMD

[a]ERN: European Reference Network.

Sustain the DCDE Implementation and Evolution

Upon receiving feedback from ERNs, the EJP RD and the EU RD Platform teams plan to offer support by organizing domain group meetings and providing templates for listing and describing DCDEs. In this sense, the EU RD Platform team formed a cancer working group, composed of cancer-related ERNs, to focus on identifying cancer DCDEs. The FAIRification stewards of the EJP RD also aim to ask all ERNs to share their data dictionaries to compare them and support the identification of commonalities within every domain group. Comparisons are meant to follow the approach and format of the existing CDD, previously performed for 4 ERNs and currently only listing CDEs. Following the conference, EJP RD also started organizing weekly web-based meetings (coffee rounds) as well as technical workshops, aiming to answer the ERNs' frequently asked questions on several topics of interest. Relevant topics included the definition and use of CDEs, CDEs minimal data set, CDEs semantic model, and modeling DCDEs. The plan is to eventually expand the semantic model of the EJP RD with identified DCDEs, publish them on the EU RD Platform alongside CDEs, and add them to ERDRI.mdr.

Discussion

Overview

This paper presented a series of joint activities aiming to extend the EU RD Platform's CDEs with DCDEs. The series starts with a strict medical phase, seeking to compile lists of DCDEs that commonly exist among registries of every ERN domain. A technical phase then follows in which a mix of medical and technical expertise primarily tackles harmonization and standardization issues. The results of each phase will be published separately. However, it is promising to review here, some of the previous efforts related to defining CDEs and identify connections to the current EU RD Platform, ERN, and EJP RD initiatives, if any.

Previous Work

The term CDEs has been first introduced to the RD field by the US initiative National Institutes of Health/National Center for Advancing Translational Sciences Global Rare Diseases Patient Registry Data Repository Program [17]. Aiming at better data standardization and interoperability for RD registries, the program defined 75 database fields required for the establishment of any RD registry [18,19]. On the basis of these attributes, the RD-Connect and EPIRARE projects developed minimum data sets for patient data entry to be used in their own framework. They also encouraged continuous alignment with the Minimal Data Elements of the European Union Committee of Experts on Rare Diseases (EUCERD) Joint Action initiative, thereby improving cooperation among RD registries at the European level [17,20-22].

Although not strictly focused on RDs, the National Institute for Neurological Disorders and Stroke initiated the Common Data Elements Project in 2005, seeking to identify the core CDEs necessary for collection in all neuroscience clinical research studies [23]. To collect CDEs, the project used case report forms (CRFs) from various clinical studies and indicated that their work was dynamic and would continue to evolve over time based on arising needs. In addition to publishing core CDEs on their website [24], they continued to identify disease-specific CDEs using a 10-step process. Similar to what has been proposed for our joint activities, their steps involved a domain working group, a draft DCDEs list, and a review process. However, deeper steps toward data standardization were also involved. In addition to defining general DCDEs for the neurological domain, they have complemented those over the years with more specific DCDEs for diseases such as epilepsy, stroke, Parkinson disease, multiple sclerosis, and headache [23]. To date, the National Institute for Neurological Disorders and Stroke CDEs project has collected data standards for 24 neurological diseases and disorders [25].

Other efforts to define DCDEs have also been made by the National Cancer Institute, which sought to identify CDEs for cancer research, thereby facilitating data interchange and interoperability between cancer research centers [20,26]. DCDEs have also been collected in a joint initiative between the Radiological Society of North America and the American College of Radiology, producing a data dictionary of radiology CDEs for various domains. These included cardiac radiology, breast imaging, chest radiology, and head and neck imaging. The initiative aimed to foster the interoperability of data present in radiologic reports and images throughout different radiologic information systems, ultimately improving research and clinical practice [27,28]. The US National Library of Medicine has also compiled a repository of >20,000 data elements, seeking to improve data quality and facilitate data comparisons among various research studies. Furthermore, it aimed to allow for opportunities to compare and combine data from multiple studies with those stored in electronic health records [20,29].

In an effort to facilitate finding necessary expertise, as well as sufficient numbers of patients for RD research, the French national minimal data set has been introduced. After systematically reviewing the scientific literature on RD CDEs, 58 data elements were represented in the data set. These were considered the clinical data standard for all French RD centers as part of the French National Plan for Rare Diseases. The methodology used to identify the minimal data set adopted the Global Rare Diseases Patient Registry Data Repository CDEs as a gold standard and also implemented many common steps with our proposed approach. These included a first working group to put together an initial CDEs draft, submitting the draft to a panel of experts, and receiving validation via a survey instrument [20,30].

Synergies With Other Activities Within the EJP RD

Our proposed approach could then be regarded as a continuation to previous efforts on DCDEs, seeking to expand the EU RD Platform's CDEs standard at the European level. It also supports ongoing and future efforts in various areas within the EJP RD. For instance, it aligns with the project's ERN-related activities, planning to hold 2 workshop series following the identification of DCDEs. The first series addresses various aspects of FAIRification, providing a set of discoverability metadata fields (metadata CDEs) that are considered the basis for describing resources and making them findable. The second series focuses

on patient matchmaking, providing a means for querying scattered patient data sets to locate similar patients with RD, either within a single ERN or across multiple ERNs. In such workshops, having DCDEs would allow the identification of discoverability metadata to be focused on certain domains. Moreover, identified DCDEs, in addition to CDEs, would present the basic parameters for queries aimed at finding similar patients. This is also the focus of the Query Builder activities of the EJP RD, running 2 pilot projects on federated discovery of resources (eg, registries and biobanks) and record-level data (eg, patients and samples). However, details of these pilots are outside the scope of this paper and will be published elsewhere.

Our approach also integrates with FAIRification activities of the EJP RD, expanding the scope of the CDEs codebook and semantic model to include DCDEs, and facilitating data exchange among institutions that use different electronic data capture software. Together with an interoperable CRF generator tool [31], the codebook content could be used by all 24 ERNs to create and reuse interoperable CRFs, sparing the need to design new electronic CRFs while implementing their registries, at least for commonly used data elements. Therefore, by incorporating DCDEs, the codebook adheres to the EU RD Platform's standard, requiring and enabling new registries to include both CDEs and DCDEs. Our efforts to define DCDEs could also take the ERN Registry TF's initiative toward a CDD further, leading to an updated version that includes DCDEs in addition to CDEs, ensuring it is harmonized among participating ERNs, and extending it to all 24 ERNs.

Conclusions

This paper presented a joint initiative of the ERNs, EU RD Platform, and EJP RD, aiming to define DCDEs for RD registration. The initiative comprises a medical and a technical phase, seeking to compile lists of DCDEs and tackle harmonization and modeling issues, respectively. Although this paper remains at a conceptual level, it starts a discussion around the importance of DCDEs and launches a series of publications presenting the methods and findings of each planned phase. From early results, based on an ERN survey and a joint conference, DCDEs seem to be an essential extension to CDEs to increase interoperability, improve discoverability, and facilitate joint research collaborations. However, at this stage, ERN registries do not seem to have clear lists of DCDEs. The approach described for defining DCDEs appears to be feasible, as it shares many common steps with previous fruitful efforts on RD CDEs, as well as with others from outside the RD field. However, it remains dynamic and should be repeated regularly by curation teams, as DCDEs are expected to evolve over time based on arising research needs.

DCDE lists will be published, alongside CDEs, on the EU RD Platform in PDF format and added to ERDRI.mdr, the technical tool serving the purpose of a data dictionary. Semantic data modeling activities of the EJP RD, which currently focus on CDEs, can also be extended to DCDEs. The number of identified domains, as well as DCDEs per domain, should remain optimally minimal, as this eases their incorporation with CDEs in all new RD registries. However, in order to avoid differences in their interpretation and implementation across ERN registries, the EU RD Platform and EJP RD both have the role of raising greater awareness and encouraging the culture change necessary for their uptake and wide use.

Acknowledgments

The description of domain-specific common data elements is supported by representatives of the European Reference Networks (ERNs): ERN-BOND, ERN CRANIO, Endo-ERN, ERN EpiCare, ERKNet, ERN-RND, ERNICA, ERN-LUNG, ERN-Skin, ERN EURACAN, ERN-EuroBloodNet, ERN eUROGEN, ERN EURO-NMD, ERN-EYE, ERN GENTURIS, ERN GUARD-HEART, ERN ITHACA, MetabERN, ERN PaedCan, ERN RARE-LIVER, ERN ReCONNECT, ERN RITA, ERN TransplantChild, and VASCERN.

The credit also goes to the FAIRification stewards of the European Joint Programme on Rare Diseases (EJP RD), César Bernabé, Shuxin Zhang, Mario Prieto, and Joeri van der Velde and the EJP RD expert, Nirupama Benis.

The work of HA, DK, BdSV, FS, and HS was supported by funding from the European Union's Horizon 2020 research and innovation program under the grant EJP RD COFUND-EJP N 825575.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of the 24 initially funded European Reference Networks.
[PDF File (Adobe PDF File), 31 KB - medinform_v10i5e32158_app1.pdf]

References

1. Taylor L. EUCERD recommendations for rare disease registries encourage interoperability. EURORDIS. 2018. URL: https://www.eurordis.org/content/eucerd-recommendations-rare-disease-registries-encourage-interoperability [accessed 2022-04-29]

2. Rare disease registries for the European Reference Networks. ECHAlliance. 2019 Jun 6. URL: https://echalliance.com/rare-disease-registries-for-the-european-reference-networks/ [accessed 2022-04-29]
3. Rare disease day: a new EU platform to support better diagnosis and treatment. European Commission. 2019 Feb 28. URL: https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1414 [accessed 2022-04-29]
4. European platform on rare disease registration (EU RD Platform). European Commission. 2019. URL: https://eu-rd-platform.jrc.ec.europa.eu/_en [accessed 2022-04-29]
5. European Joint Programme on Rare Diseases. 2019. URL: https://www.ejprarediseases.org/ [accessed 2022-04-29]
6. European joint programme on rare diseases (EJP RD) General Assembly. eUROGEN, European Reference Network. 2019 Oct 1. URL: https://eurogen-ern.eu/european-joint-programme-rare-diseases-ejp-rd-general-assembly/ [accessed 2022-04-29]
7. European rare disease registry infrastructure (ERDRI). European Commission. 2019. URL: https://eu-rd-platform.jrc.ec.europa.eu/erdri-description_en [accessed 2022-04-29]
8. Set of common data elements for rare disease registration. European Commission Joint Research Center. 2019. URL: https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/CDS/EU_RD_Platform_CDS_Final.pdf [accessed 2022-04-29]
9. Set of common data elements. European Commission. 2019. URL: https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en [accessed 2022-04-29]
10. Common data element (CDE) resource portal: glossary. National Library of Medicine. 2013. URL: https://tinyurl.com/mw3vjvvj [accessed 2022-04-29]
11. 1st General Assembly of the European joint programme on rare disease research. ERN Transplant-Child. 2019. URL: https://us16.campaign-archive.com/?u=8e7edf118f0e6c9deea7a292f&id=7ac0e0eb48 [accessed 2022-04-29]
12. In its first workshop, the ERN Research Working Group defined what ERN research entails. European Reference Networks. 2019. URL: https://ec.europa.eu/health/latest-updates/its-first-workshop-ern-research-working-group-defined-what-ern-research-entails-2019-03-28_en [accessed 2022-04-29]
13. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
14. Warner J, Johnston M, Korngut L, Jette N, Pringsheim T. Common data elements for neurological registries. Can J Neurol Sci 2013 Jul;40(4 Suppl 2):S62-S63. [doi: 10.1017/s0317167100017212] [Medline: 23787272]
15. VASCA common data elements (CDE) - datasets. Amsterdam UMC. 2020. URL: https://decor.nictiz.nl/art-decor/decor-datasets--vasca-?id=&effectiveDate=&conceptId=&conceptEffectiveDate= [accessed 2022-04-29]
16. Semantic data model of the set of common data elements for rare disease registration. GitHub. 2020. URL: https://github.com/ejp-rd-vp/ERN-common-data-elements/wiki [accessed 2022-04-29]
17. Registry common data elements (CDEs). RD-Connect. 2018. URL: https://rd-connect.eu/what-we-do/phenotypic-data/registry-common-data-elements/ [accessed 2022-04-29]
18. Rubinstein YR, McInnes P. NIH/NCATS/GRDR® Common data elements: a leading force for standardized data collection. Contemp Clin Trials 2015 May;42:78-80 [FREE Full text] [doi: 10.1016/j.cct.2015.03.003] [Medline: 25797358]
19. Office of rare diseases research global rare diseases patient registry and data repository-GRDR minimal common data elements (CDEs). National Institutes of Health. URL: https://rarediseases.info.nih.gov/files/list_cdes.pdf [accessed 2022-04-29]
20. Kodra Y, Weinbach J, Posada-de-la-Paz M, Coi A, Lemonnier SL, van Enckevort D, et al. Recommendations for improving the quality of rare disease registries. Int J Environ Res Public Health 2018 Aug 03;15(8):1644 [FREE Full text] [doi: 10.3390/ijerph15081644] [Medline: 30081484]
21. Taruscio D, Mollo E, Gainotti S, Posada de la Paz M, Bianchi F, Vittozzi L. The EPIRARE proposal of a set of indicators and common data elements for the European platform for rare disease registration. Arch Public Health 2014 Oct 13;72(1):35 [FREE Full text] [doi: 10.1186/2049-3258-72-35] [Medline: 25352985]
22. EUCERD core recommendations on rare disease patient registration and data collection to the European Commission, member states and all stakeholders. EUCERD - European Union Committee of Experts on Rare Diseases. 2013 Jun 5. URL: https://rarediseases.org/wp-content/uploads/2015/12/EUCERD_Recommendations_RDRegistryDataCollection_adopted.pdf [accessed 2022-04-29]
23. Grinnon ST, Miller K, Marler JR, Lu Y, Stout A, Odenkirchen J, et al. National Institute of Neurological Disorders and Stroke common data element project - approach and methods. Clin Trials 2012 Jun;9(3):322-329 [FREE Full text] [doi: 10.1177/1740774512438980] [Medline: 22371630]
24. NINDS common data elements - Harmonizing information. Streamlining research. National Institute of Neurological Disorders and Stroke. URL: https://www.commondataelements.ninds.nih.gov/ [accessed 2022-04-29] [WebCite Cache ID 6DtwForbc]
25. Suarez JI, Sheikh MK, Macdonald RL, Amin-Hanjani S, Brown Jr RD, de Oliveira Manoel AL, Unruptured Intracranial Aneurysms and SAH CDE Project Investigators. Common data elements for unruptured intracranial aneurysms and subarachnoid hemorrhage clinical research: a national institute for neurological disorders and stroke and national library of medicine project. Neurocrit Care 2019 Jun;30(Suppl 1):4-19. [doi: 10.1007/s12028-019-00723-6] [Medline: 31087257]

26. Nadkarni PM, Brandt CA. The common data elements for cancer research: remarks on functions and structure. Methods Inf Med 2006;45(6):594-601 [FREE Full text] [Medline: 17149500]
27. Rubin DL, Kahn Jr CE. Common data elements in radiology. Radiology 2017 Jun;283(3):837-844. [doi: 10.1148/radiol.2016161553] [Medline: 27831831]
28. Common data elements (CDEs) for radiology. Radiological Society of North America. URL: https://www.radelement.org/ [accessed 2022-04-29]
29. Common data element (CDE) resource portal. National Library of Medicine. URL: https://cde.nlm.nih.gov/home [accessed 2022-04-29]
30. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J Am Med Inform Assoc 2015 Jan;22(1):76-85 [FREE Full text] [doi: 10.1136/amiajnl-2014-002794] [Medline: 25038198]
31. de Ridder S. The iCRF generator. GitHub. 2019. URL: https://github.com/aderidder/iCRFGenerator/ [accessed 2022-04-29]

Abbreviations

CDD: Common Data Dictionary
CDE: common data element
CRF: case report form
DCDE: domain-specific common data element
EJP RD: European Joint Programme on Rare Diseases
EPIRARE: European Platform for Rare Disease Registries
ERDRI: European Rare Disease Registry Infrastructure
ERN: European Reference Network
EU RD Platform: European Platform on Rare Disease Registration
EUCERD: European Union Committee of Experts on Rare Diseases Joint Action
JRC: Joint Research Centre
RD: rare disease
TF: task force

Edited by C Lovis; submitted 16.07.21; peer-reviewed by T Fawzi; comments to author 03.10.21; revised version received 28.11.21; accepted 02.01.22; published 20.05.22.

Please cite as:
Abaza H, Kadioglu D, Martin S, Papadopoulou A, dos Santos Vieira B, Schaefer F, Storf H
Domain-Specific Common Data Elements for Rare Disease Registration: Conceptual Approach of a European Joint Initiative Toward Semantic Interoperability in Rare Disease Research
JMIR Med Inform 2022;10(5):e32158
URL: https://medinform.jmir.org/2022/5/e32158
doi:10.2196/32158
PMID:35594066

Original Paper

Construction of a Linked Data Set of COVID-19 Knowledge Graphs: Development and Applications

Haofen Wang[1], PhD; Huifang Du[1], MSc; Guilin Qi[2], PhD; Huajun Chen[3], PhD; Wei Hu[4], PhD; Zhuo Chen[3], BSc

[1]College of Design and Innovation, Tongji University, Shanghai, China

[2]School of Computer Science and Engineering, Southeast University, Nanjing, China

[3]College of Computer Science and Technology, Zhejiang University, Hangzhou, China

[4]National Institute of Healthcare Data Science, Nanjing University, Nanjing, China

Corresponding Author:
Haofen Wang, PhD
College of Design and Innovation
Tongji University
No. 281 Fuxin Road, Yangpu District
Shanghai, 200092
China
Phone: 86 13918586855
Email: carter.whfcarter@gmail.com

Abstract

Background: With the continuous spread of COVID-19, information about the worldwide pandemic is exploding. Therefore, it is necessary and significant to organize such a large amount of information. As the key branch of artificial intelligence, a knowledge graph (KG) is helpful to structure, reason, and understand data.

Objective: To improve the utilization value of the information and effectively aid researchers to combat COVID-19, we have constructed and successively released a unified linked data set named OpenKG-COVID19, which is one of the largest existing KGs related to COVID-19. OpenKG-COVID19 includes 10 interlinked COVID-19 subgraphs covering the topics of encyclopedia, concept, medical, research, event, health, epidemiology, goods, prevention, and character.

Methods: In this paper, we introduce the key techniques exploited in building COVID-19 KGs in a top-down manner. First, the schema of the modeling process for each KG in OpenKG-COVID19 is described. Second, we propose different methods for extracting knowledge from open government sites, professional texts, public domain–specific sources, and public encyclopedia sites. The curated 10 COVID-19 KGs are further linked together at both the schema and data levels. In addition, we present the naming convention for OpenKG-COVID19.

Results: OpenKG-COVID19 has more than 2572 concepts, 329,600 entities, 513 properties, and 2,687,329 facts, and the data set will be updated continuously. Each COVID-19 KG was evaluated, and the average precision was found to be above 93%. We have developed search and browse interfaces and a SPARQL endpoint to improve user access. Possible intelligent applications based on OpenKG-COVID19 for further development are also described.

Conclusions: A KG is useful for intelligent question-answering, semantic searches, recommendation systems, visualization analysis, and decision-making support. Research related to COVID-19, biomedicine, and many other communities can benefit from OpenKG-COVID19. Furthermore, the 10 KGs will be continuously updated to ensure that the public will have access to sufficient and up-to-date knowledge.

(JMIR Med Inform 2022;10(5):e37215) doi:10.2196/37215

KEYWORDS

knowledge graph; linked data; COVID-19; knowledge extraction; knowledge fusion; natural language processing; artificial intelligence; data set; schema modeling; semantic search

Introduction

On February 11, 2020, the World Health Organization announced the official name of the 2019 novel coronavirus as COVID-19. Meanwhile, the International Committee on Taxonomy of Viruses named this novel coronavirus SARS-CoV-2 [1]. The infection caused by SARS-CoV-2 is now affecting almost every country in the world. By October 24, 2021, more than 4.95 million people have died from COVID-19, raising concerns of widespread fear and increasing anxiety in individuals. At present, the epidemic continues to spread, and there are many questions that continue to plague the public about this disease, including: How can we obtain an overall understanding of the knowledge about COVID-19 facing such large amounts of information coming from various media every day? What are the variants of SARS-CoV-2 and how should they be treated or prevented? What is the state of supplies, hot events, and frontline health care workers in this invisible war worldwide? How can we find drugs or vaccines, and further learn more? What travel restrictions do local policies apply during the epidemic? What are the requirements regarding the various means of transport?

During this pandemic, artificial intelligence (AI) has served as an enabler to combat COVID-19, such as successful attempts in predicting epidemic trends [2] with sophisticated models, accelerating computer tomography detection [3] for more efficient diagnosis by computer vision, participating in drug development [4], and automatically answering epidemic-related natural language questions [5-7]. Besides deep learning, the knowledge graph (KG) concept has drawn increasing attention from both academia and industry since it was first proposed by Google in 2012. As the key to the evolution of AI toward cognitive intelligence, a KG enables machines to better organize, reason, understand, and explain information.

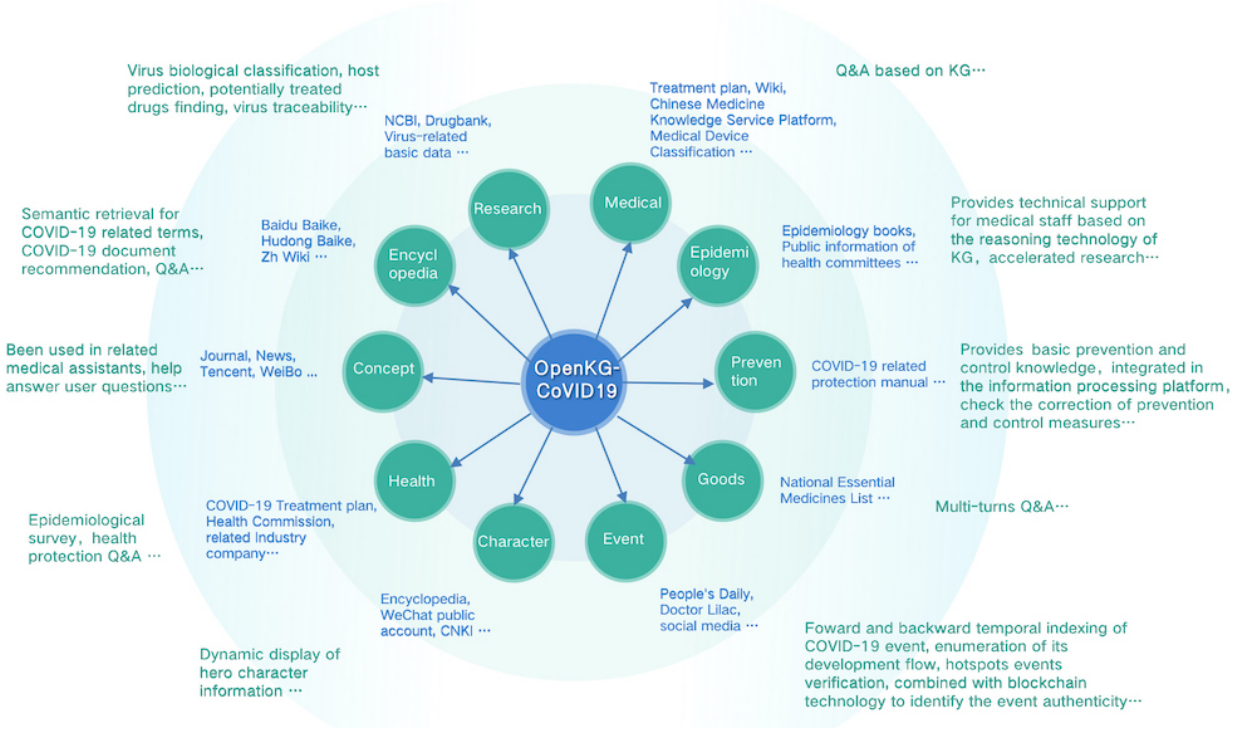
The success of the above applications heavily depends on the scale and quality of the underlying KGs, regardless of whether they exist in the open or in a specific domain. Well-known general-purpose KGs include DBpedia [8], Yago [9], Freebase [10], Wikidata [11], and the Chinese linked open data effort Zhishi.me [12]. All of these KGs leverage Wikipedia, one of the largest encyclopedia websites in the world, as an important source. WordNet [13], BabelNet [14], and Linguistic Linked Open Data [15] are examples of linguistic KGs. Regarding domain-specific KGs, we here mainly focus on life science or health care fields. The KG Linking Open Drug Data [16] surveys the publicly available data about drugs and creates linked representations of the data sets. The project Open PHACTS [17] aims to deliver and sustain an open pharmacological space using and enhancing state-of-the-art semantic web standards and technologies. Bio2RDF [18] uses semantic web technologies to provide the largest network of linked data about the life sciences. However, none of the above KGs is specific to COVID-19. Although it is possible to extract a COVID-19–relevant subgraph from general-purpose KGs, this approach will suffer from low coverage of domain knowledge and the sparsity of properties describing this knowledge (eg, viruses and diseases).

The White House, in collaboration with publishers and tech firms, has launched the CORD-19 data set [19], which contains more than 59,000 published articles and preprints. Although CORD-19 is considered to be the largest single collection of COVID-19 knowledge amassed to date, the majority of the data set contains unstructured data, and more than 60% of the included papers do not mention search terms such as “coronavirus” and “SARS-CoV” [20]. The existing COVID-19 Knowledge Graph [21] is an expansive cause-and-effect network constructed from the scientific literature on SARS-CoV-2, aiming to provide a comprehensive view of its pathophysiology. However, there are only 10 entity types and 9484 facts within this KG. Coronavirus Knowledge Graph [22] only has 27 relation types. The CovidGraph project [23] built a COVID-19 graph that stores publications, case statistics, and molecular data in a Neo4j database, which enables exploring the underlying knowledge for finding specific genes, authors, articles, patents, proteins, existing treatments, and medications relevant to the entire family of coronaviruses. However, key aspects such as health care, epidemiology, antiepidemic goods, related events, and frontline workers fighting the epidemic have not yet been considered.

To capture richer and more diverse topics of COVID-19 so as to offer more useful knowledge for the public, we have extended these previous efforts [24-26] to construct OpenKG-COVID19, a linked data set of COVID-19 KGs, covering 10 aspects ranging from encyclopedia, concept, medical, health, prevention, goods, research, epidemiology, and character to events. OpenKG-COVID19 was launched by OpenKG [27], which is the largest Chinese open KG community pushing for the development of public KGs, open-source tools, and best practices in vertical sectors in China since the middle of February 2020. We are the first to mainly focus on constructing high-quality pandemic KGs in China. Moreover, OpenKG-COVID19 is open to the public with continuous efforts to ensure that it contains up-to-date information. The publishing and maintenance of such a large-scale KG can help researchers around the world to understand, study, and even fight COVID-19. An overview of OpenKG-COVID19 is depicted in Figure 1. Each KG, its sources, and possible applications are listed in Textbox 1.

Moreover, several key steps have been used to construct OpenKG-COVID19, namely modeling, extraction, and fusion of knowledge. Among them, knowledge modeling mainly involves schema design. The schema knowledge of each data set in OpenKG-COVID19 is described in the Methods section. The other steps are executed automatically with the human in the loop. In particular, we present the technical details of knowledge extraction and then describe how the curated KGs are further linked together at both the schema and data levels. We further present the results of experimental validation of OpenKG-COVID19, and discuss the access interfaces along with the possible applications of the linked COVID19 KGs.

Figure 1. Overview of OpenKG-COVID19. KG: knowledge graph; NCBI: National Center for Biotechnology Information; Q&A: question and answer.



Textbox 1. Sources, knowledge graphs (KGs), and application prospects of OpenKG-COVID19.

- The encyclopedia KG (Bilingual Encyclopedia Knowledge Graph [BEKG]) is based on multiple encyclopedia sources, which helps to gain a basic understanding of SARS-CoV-2 and COVID-19.
- Targeting the question and answer (QA) system, both the medical KG and the health KG consider data sources from industrial companies and official treatment plans, which have included COVID-19–related symptoms, diseases, drugs, and treatment options.
- The prevention KG not only provides authoritative guidance on individuals' protection and public prevention, but also contains knowledge about vaccines and nucleic acid tests.
- The goods KG provides the current status of materials used in the epidemic, including information of daily protective equipment, medical diagnosis, treatment devices, and therapeutic drugs.
- The research KG aims to assist in the discovery of drugs or vaccines, and its data are derived from virus-related scientific research databases and literature.
- The epidemiology KG helps to trace the source of infection and explore contacts. These data come from the case flow information published by provincial health committees.
- The character KG sorts out heroic deeds and assists in the dynamic display of character information, including the individual's resume, achievements, and related events about combating the epidemic.
- The event KG organizes hot events about the epidemic with the when, where, who, and what factors incorporated.
- The concept KG uses automatic web-mining technologies to collect a large number of fine-grained COVID-19–related entities and their corresponding hypernyms from web text, which has been applied in medical-related virtual assistants to address complex user information needs.

Methods

Schema of OpenKG-COVID19

A schema defines a specific, clear, high-level structure of a KG. It is necessary to model a sound schema to accurately offer a clear understanding of KG content. New data added to the KG will not be allowed if the data do not conform to the defined schema. We designed a total of 10 schemata for each subgraph: concept, encyclopedia, medical, health, research, prevention, goods, event, character, and epidemiology. The details of the schemata are described in further detail elsewhere [28]. In brief, three methods were employed to develop the schemata: manually defined by medical experts (manual), extracted from encyclopedic websites or COVID-19–related medical websites (site data), and mined automatically from the web (automatic mining). The design method of each KG is displayed in the left part of Table 1.

Within OpenKG-COVID19, the schemata of most KGs (eg, medical, epidemiology) have been designed by domain experts. Taking the epidemiology KG as an example, its schema defines the basic concepts of epidemiology such as epidemic, pathogen, host, epidemic situation, epidemiological survey, survey method, survey population, surveyed individual, and survey report. The relations between these concepts contain "cause," "is-part-of,"

"includes," "uses," and similar. The entire schema diagram is illustrated in Figure 2. Note that even though the schema shown here was manually constructed, we can boost the entire process by recommending users' domain keywords or related ontologies in the same or similar field as a prototype for reuse.

Another method for schemata design is to treat semistructured information as categories and properties in "infoboxes" as schemata. This method was used for the design of the schemata for the encyclopedia and research KGs. Specifically, the schema modeling process during construction of the encyclopedia KG is shown with a red color border in Figure 3. We further used BabelNet [29] and Zhishi.schema [30] to expand the concepts with multilingual labels.

We also tried to automatically mine schemata from the web. Specifically, we performed nonlinear mapping between one concept to another (its hypernym) based on popular embedding technology to obtain a large number of fine-grained hypernyms from search engines, encyclopedias, and word morphology. The hierarchical structure ("is-a" relation) was constructed by measuring the semantic broadness between concepts as well as between an instance and a concept. Therefore, the data-level knowledge was also extracted during schema design.

Table 1. Classifications of schema design and knowledge extraction of COVID-19 knowledge graphs.

Knowledge graph	Schema design			Knowledge extraction		
	Manual	Site data	Automatic mining	Structured	Semistructured	Plain text
Concept			✓		✓	✓
Encyclopedia		✓	✓		✓	✓
Medical	✓					✓
Health	✓	✓			✓	
Research		✓		✓		✓
Prevention	✓			✓	✓	✓
Goods		✓			✓	
Event	✓					✓
Character		✓			✓	
Epidemiology	✓			✓		

Figure 2. Schema diagram of the epidemiology knowledge graph.

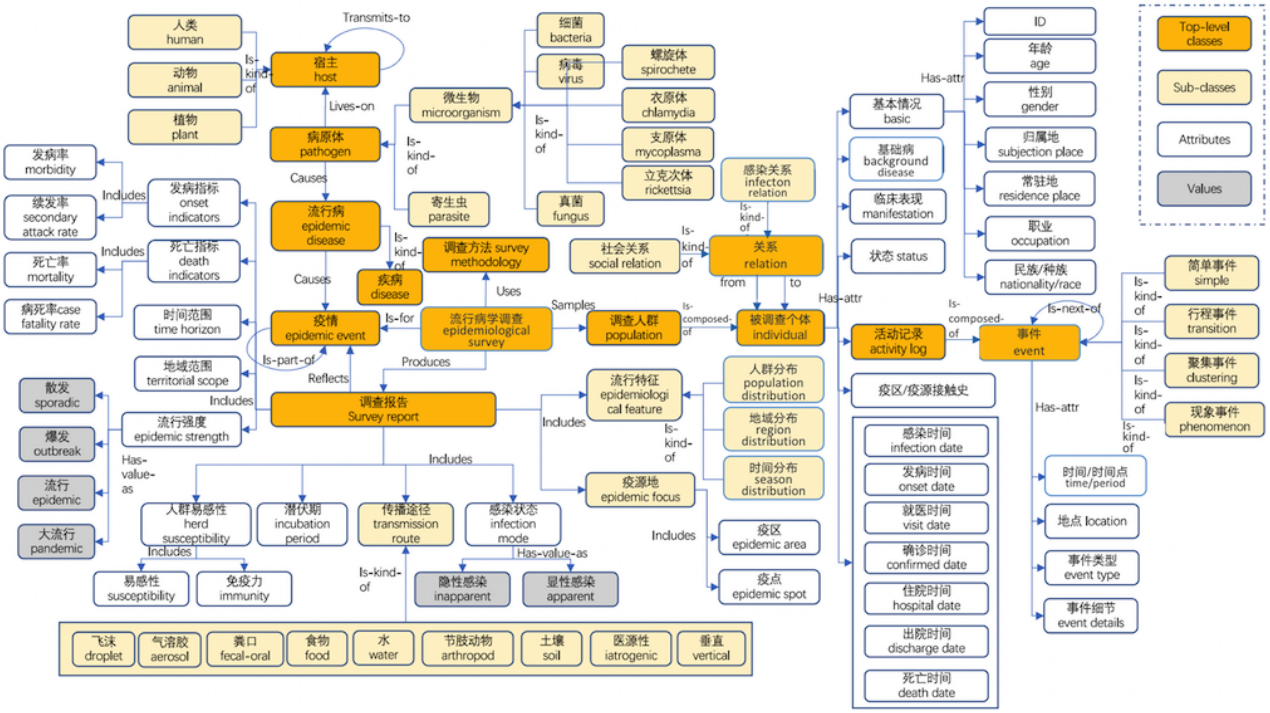
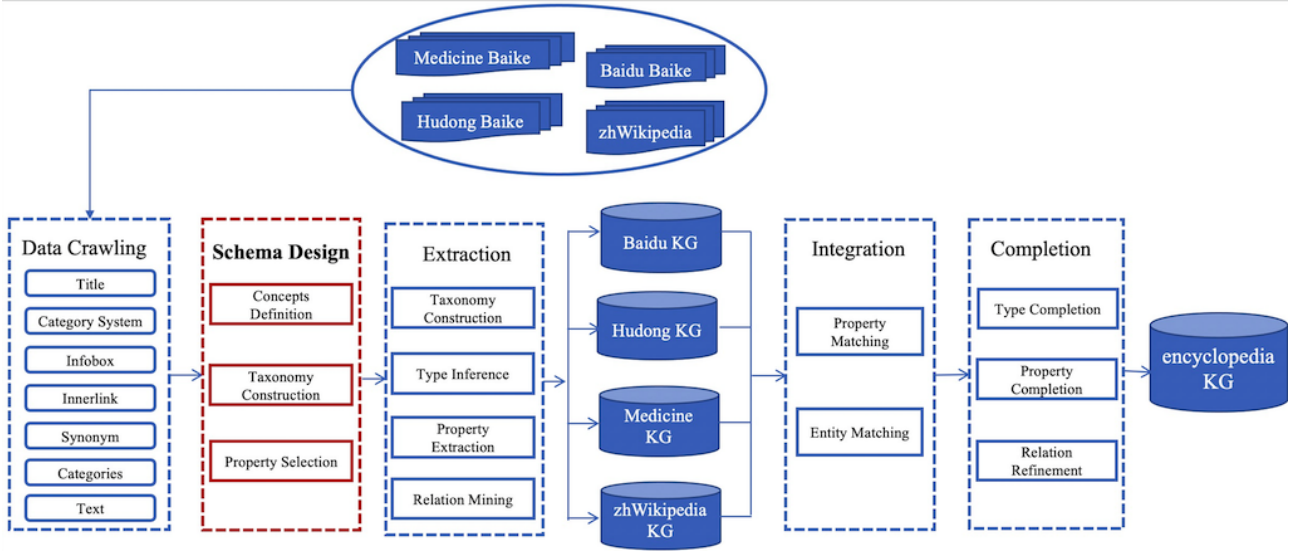


Figure 3. Construction process of the encyclopedia knowledge graph (KG).



Knowledge Extraction for COVID-19 KGs Construction

Overview

This section introduces methods for the classification of knowledge extraction based on different data types. In general, sources of knowledge extraction include structured data (eg, linked data), semistructured data (eg, tables and infoboxes), and unstructured data in the form of plain text. Sources of each KG in OpenKG-COVID19 are listed in the right part of Table 1. There exist correlations between schema design and sources for knowledge extraction. For example, if a KG extracts its knowledge from semistructured data sources, its schema is usually obtained from site data. Graph mapping is leveraged to extract a domain-specific subgraph from linked data, whereas the "D2R" tool is used to transform relational data of a Database into Resource Description Framework (RDF) triples. Moreover, wrappers are used for semistructured data and information extraction to convert plain text into structured knowledge.

Extraction from Structured Data Sources

Structured data represent the main data source of KG construction. Our research KG focuses on information from the virus field. It contains five subgraphs, which are the virus taxonomy KG, SARS-CoV-2 gene-protein KG, antiviral drug KG, SARS-CoV-2 phylogeny KG, and SARS-CoV-2 literature extraction KG. The construction of the first four KGs fits within this method.

Specifically, we analyzed some data of related biodatabases (eg, National Center for Biotechnology Information [NCBI] [31], GISAID [32], China National Center for Bioinformation [33], DrugBank [34], and Nextstrain [35]) and related biological KGs such as SNAP [36] at Stanford University. Moreover, we have established in-depth collaborations with some biological institutes in the vertical field to ensure that the research KG is professional. We converted data in different formats from the above sources into a unified graph structure based on the designed schema.

The SARS-CoV-2 gene-protein KG is mainly built from the virus data in the NCBI database. By looking up "SARS-CoV-2" in NCBI, various types of related information are returned, such as genome, gene, and protein. Two example triples are (SARS-CoV-2, Virus-express-Gene, NS6) and (SARS-CoV-2, Virus-produce-Protein, nonstructural protein NS6).

The antiviral drug KG is based on four structured databases: DrugBank, Virus Pathogen Database [37], VirHostNet 3.0 [38], and VISDB [39]. The KG demonstrates interaction relationships among various types of viruses, human proteins, antiviral drugs, and diseases. For further integration, we linked the data through the taxonomy ID of the virus, the UniProt ID of the protein, and the generic name of the drug. Several extracted example triples are: (Human immunodeficiency virus 1, Virus-alias-String, HIV-1), (Enfuvirtide, Drug-effect-Virus, Human immunodeficiency virus 1), and (H31, HostProtein-belong to-Host, Human).

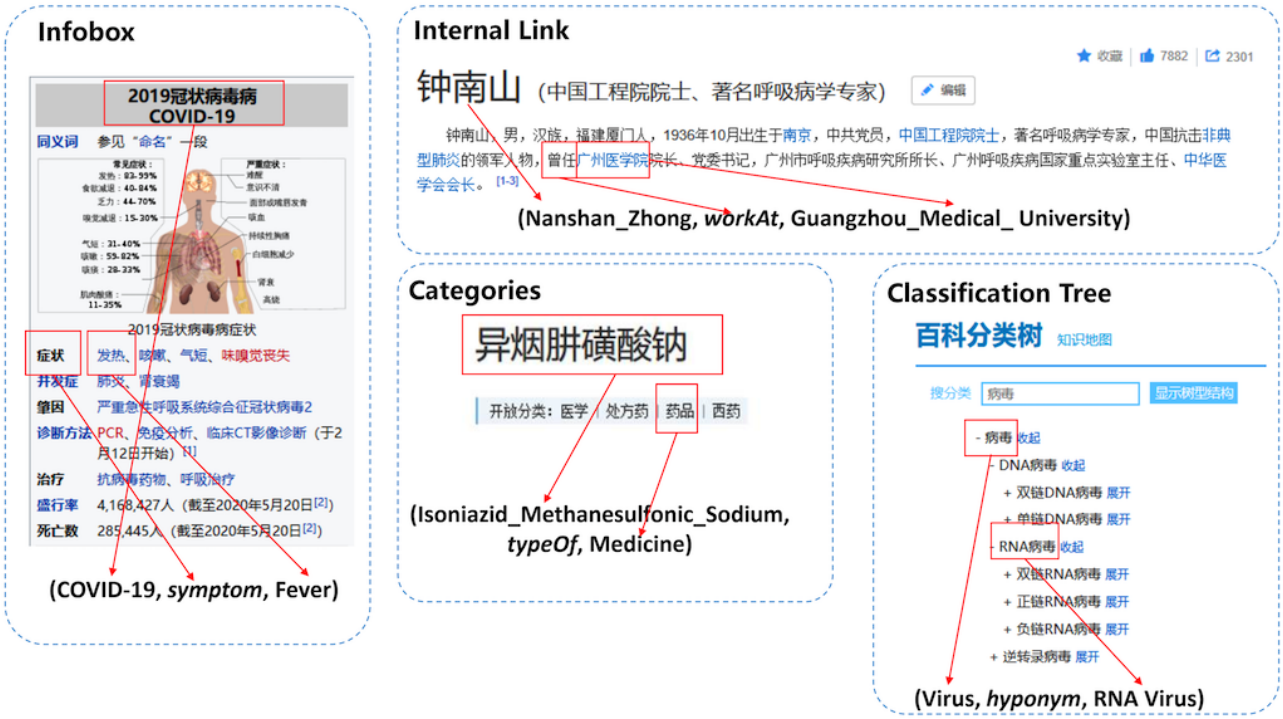
We also extracted the virus taxonomy tree from NCBI to build the corresponding KG. Similarly, the SARS-CoV-2 phylogeny KG was constructed by referencing Nextstrain metadata.

Extraction From Semistructured Sites

We mainly leveraged semistructured data for building the KGs of concept, encyclopedia, health, prevention, goods, and character. Taking the encyclopedia KG as an example, its knowledge in the form of RDF triples is extracted from the integration of several encyclopedia sites (eg, Baidu Baike, Hudong Baike, Chinese Wikipedia). We particularly considered the following four types of semistructured data for knowledge extraction: internal links, infoboxes, categories, and classification trees. For an infobox, the page title is treated as a subject, each attribute of the infobox is treated as a predicate, and the corresponding attribute value is treated as the object. For an internal link, we also treat the title entity as a subject, the target entity that the internal link refers to as the object, and the relation (defined in the schema) matching the text between the subject mention and the object mention as the predicate. For a category that a page belongs to, the title entity, typeOf, and the given category form a triple. For a classification tree, a

high-level concept can be linked with any of its ancestors in a triple using the hyponym as the predicate. Figure 4 shows examples for all four types of semistructured data and their corresponding extracted triples.

Figure 4. Extraction of various types of semistructured data.



Extraction From Plain Texts

Plain texts are widely available for human consumption but are hard for machines to understand, which hinders construction of a KG from these unstructured data. We applied regular expressions to extract fact triples for the six subgraphs shown in the right part of Table 1 from plain texts. When detecting knowledge by regular expressions, we paid more attention to the precision of information extraction rather than the recall to ensure that the COVID-19 knowledge managed into OpenKG-COVID19 is relatively accurate. Finally, the average precision of our regex matching methods was found to be 96.34% and the average recall was 87.63%. However, there are large amounts of diverse information and complex semantic relations in the research literature, which required more advanced methods during the construction of the research KG. In recent years, there has been great progress in applying machine reading comprehension to the knowledge extraction task on plain texts [40,41]. The basic idea is to extract the candidate entities from sentences by a subject extraction network, and then extract the object of a triple based on candidate entities and a predefined predicate using a joint predicate-object extraction network. Pretrained language models such as bidirectional encoder representations from transformers (BERT) [42] are employed for encoding in both networks, which alleviates the amount of labeled data required to train a model.

Inspired by the above work, we applied the same technique in building COVID-19 KGs from various text sources. The labeling process can be further relieved by distant supervision [43], where the subject and object of a triple are automatically labeled in one sentence and the sentence context is captured to check whether the predicate holds. After extraction and sampled manual check, triples such as (SARS-CoV2, Virus-interaction-Human Protein, ACE2), (SARSCoV-2, Virus-cause-Disease, human respiratory disease), and (nelfinavir, Drug-effect-Virus, SARS-CoV2) are returned from the medical literature.

Interlinking Knowledge from Different COVID-19 KGs

Overview

Following the linked data principles, we connected these KGs to promote the integration and sharing of knowledge about COVID-19. We observed that schemata in these KGs, except for that of the concept KG, are of relatively small scale. Therefore, we first used an automatic ontology matching approach to align schema-level knowledge (ie, concepts and properties) and then asked domain experts to validate the results, and finally leveraged the validated schema matches to align data-level knowledge (ie, entities).

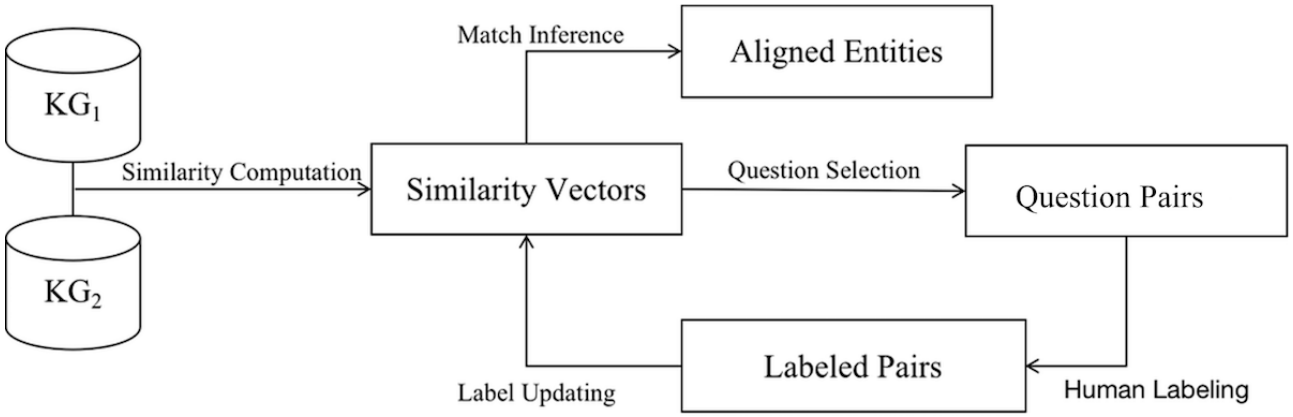
Schema Matching

Because there is no central schema for the COVID-19 KGs, we decided to conduct pairwise schema matching. We reused Falcon-AO [44], which is an automatic ontology matching tool. Its main strengths lie in the integration of various powerful matchers exploiting linguistic and structural features. Furthermore, due to the naming issue, many schemata use sequential IDs to name their concepts and properties. To avoid their interference with the matching process, we disabled the comparison of local names in Falcon-AO. The details of the naming convention are introduced below.

Entity Alignment

Similar to schema matching, we conducted pairwise entity alignment. We used the property matches from schema matching to make the properties in each pair of KGs uniform. We observed that most matched properties are data-type properties. Therefore, we leveraged literal similarity measures to align entities. Since the true matches are unavailable, we deployed the crowdsourced entity resolution approach [45] to find entity matches, and the workflow is depicted in Figure 5.

Figure 5. Workflow of entity alignment.KG: Knowledge graph.



Similarity Computation

For each entity pair, we used similarity measures to construct a similarity vector, where each real value in the vector represents the similarity of the values of each pair of aligned properties. For numerical values, the absolute difference similarity measure was used. For textual values, the Jaccard similarity measure was applied. Moreover, the entity-type values were converted to texts based on their labels. Note that a few KGs are multilingual; therefore, we used a character-level bigram to tokenize textual values.

Match Inference

Based on similarity vectors of entity pairs, we used the partial order assumption to infer matches and nonmatches. Once an entity pair is judged as a match by a human, each entity pair such that all similarity values are not less than those of the match is inferred as a match. By contrast, once an entity pair is judged as a nonmatch, each entity pair such that all similarity values are not greater than those of the nonmatch is inferred as a nonmatch. When the similarity measures evaluate the value within the threshold range, these inference rules are approximately true [46].

Question Selection

To save both human labor and time, the total number of questions (ie, unresolved entity pair for validation) is required to be minimized. However, the true answers for questions are unknown. Alternatively, we maximized the inference power of a new question in each step. The question-selection algorithm iteratively chooses each unresolved pair that has the greatest number of possible inferred matches and nonmatches.

Human Labeling

Some KGs contain a lot of medical details (eg, drugs in research, posthospital medications, limitations, special diets); thus, common workers from the crowdsourcing platforms may not have sufficient domain knowledge to manage a large amount of medical information. To ensure a data set of high quality and benefit to downstream tasks such as question answering, we employed expert sourcing instead of crowdsourcing to collect answers for questions pairs. In detail, we asked one domain expert to judge each unresolved pair as a match or a nonmatch, and randomly sampled some labeled question pairs for further review to obtain the final result.

Results

Data Evaluation

Data Statistics

OpenKG-COVID19 is a linked data set of COVID-19 KGs consisting of 10 subgraphs derived from different sources such as research publications, medical guidelines, and encyclopedia websites. As of October 24, 2021, the data set has knowledge of more than 2572 concepts, 329,600 entities, 513 properties, and 2,687,329 facts. Moreover, the data set will be updated continuously along with the occurrence of COVID-19. The detailed statistics of each KG are listed in the left part of Table 2, demonstrating that the research KG contains the largest numbers of both entities and facts, and all KGs have relatively rich properties, except for the concept KG that only defines two properties (ie, typeOf and subClassOf) but has the highest number of concepts.

Table 2. Detailed statistics and quality of each subgraph.

Knowledge graph	Facts, n	Concepts, n	Entities, n	Properties, n	Evaluation, n	Correct, n	Precision (%), mean (SD)
Encyclopedia	261,154	50	54,318	60	5000	4778	95.52 (0.58)
Medical	2857	54	1035	92	652	620	94.81 (1.71)
Research	2,281,797	31	221,131	64	8556	8555	99.96 (0.03)
Event	27,388	4	2291	21	200	198	96.35 (1.69)
Character	1902	21	1057	40	570	570	99.65 (0.35)
Prevention	28,651	113	34,859	24	646	630	97.20 (1.25)
Goods	3738	165	132	57	365	359	97.83 (1.42)
Health	51,575	592	7110	104	487	483	98.78 (0.91)
Epidemiology	8336	55	2163	47	200	200	98.08 (1.92)
Concept	19,391	1487	4784	2	100	96	92.31 (4.96)

Accuracy Evaluation

It is crucial to assess the quality of each KG in OpenKG-COVID19. Since no ground truths are available, we performed manual evaluation. Owing to the large number of facts, we adopted a similar method as that of Yago with respect to the sampling strategy and labeling process.

For sampling, we evaluated a chosen sample of facts for each property defined in OpenKG-COVID19. Since the fact number of each property is not evenly distributed, we used different sampling coefficients (ranging from 0 to 1) for different properties. If the fact number of one property is lower than the minimal sample number k (k=20 in our setting), it was set to 1. Otherwise, we selected a random coefficient to ensure that the returned samples are more than k.

For labeling, we invited three postgraduate students focusing on KGs as their main research area to review the same sampling data for each subgraph. They were offered three choices to annotate each sample: agree, disagree, and unknown. If more than one annotator made a certain choice, then the sample was labeled as that choice. If there were three different annotations for one sample, we asked the annotators to reconsider the choice through acquiring further knowledge about the sample and obtain a result. However, discrepancies only accounted for 6% of all samples according to the record of the labeling process. After the labeling process, 98.35% of the sampled facts were considered to be correct by consensus. To generalize our results on the subset to the whole data set of COVID-19 KGs, the Wilson interval at α=5% was computed.

The precision value of each COVID-19 KG is reported in the right part of Table 2. We found that all KGs achieved an average precision of more than 93%, except for the concept KG with knowledge extracted by automatic web mining, which indicates the high quality of OpenKG-COVID19. After the error analysis, we found two typical patterns of wrong facts. One is that there exists a mistake of either the head entity or the tail entity, and the other is that the relation between the entity pair does not conform to the fact. For example, it is inappropriate to regard "judgment basis" as the relation between "confirmed cases" and "shock," because this is simply a possible clinical manifestation of patients with COVID-19.

Results and Quality of Interlinking

The schema matching results are shown in Table 3, demonstrating overlaps between different schemata, although such overlaps are limited. Regarding entity alignment, we found 1055 matches among five KGs. The encyclopedia KG had the greatest number of matches with other KGs (ie, 836 with the health KG, 55 with the medical KG, 11 with the character KG, and 2 with the goods KG) because it contains various types of entities (eg, drugs and hospitals). We also noted some entity matches but no schema matches between the encyclopedia KG and the character KG, because some shared properties (eg, rdfs:label) are used to align entities but these properties are not included in schema matching. We also found some duplicated entities in the encyclopedia KG because these entities are extracted from different websites. There were few matches between the goods KG and other KGs because most entities in the goods KG are medical devices, which do not appear in the other KGs. Since some entities in the character KG are hospitals, there were 19 matches with the health KG. The remaining matches were mostly related to drugs.

We recruited three students with a major in Semantic Web to evaluate the precision, recall, and F1-score of the schema matching and entity alignment results. As shown in Table 4, the schema matching achieved high recall, but relatively low precision. Most false matches were caused by the similarity measure (eg, the pair "determination of protein" and "protein" was wrongly judged as a match). We observed that the entity alignment achieved perfect results in all KG pairs except for health-character with precision, recall, and F1-score of 88.2%, 100.0%, and 93.8%, respectively. The high performance of entity alignment was attributed to the fact that the literal information in KGs is of high quality and most matches share exactly the same information.

Table 3. Results of schema matching.[a]

Knowledge graph	Encyclopedia	Prevention	Concept	Health	Research	Medical	Epidemiology	Event	Goods	Character
Encyclopedia	—[b]	0	0	9	4	6	1	0	0	0
Prevention	0	—	0	0	0	2	0	16	17	1
Concept	37	7	—	0	0	0	0	0	0	0
Health	9	0	41	—	3	13	3	1	1	0
Research	2	0	6	2	—	3	0	1	0	0
Medical	4	2	25	4	4	—	6	3	0	1
Epidemiology	4	0	18	3	0	3	—	0	0	4
Event	0	5	1	0	0	0	1	—	16	0
Goods	0	1	18	6	0	0	0	2	—	0
Character	0	2	11	0	5	5	3	1	0	—

[a]The numbers below the diagonal are class matches and the numbers above the diagonal are property matches.

[b]Not applicable.

Table 4. Performance of schema matching.

Knowledge graph	Class (%)			Property (%)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Encyclopedia	75.0	85.7	80.0	85.0	85.0	85.0
Prevention	76.5	100.0	86.7	94.4	100.0	97.1
Concept	72.0	90.1	80.0	—[a]	—	—
Health	55.4	94.7	69.9	76.7	88.5	82.1
Research	78.9	100.0	88.2	54.5	100.0	70.6
Medical	87.2	91.1	89.1	79.4	90.0	84.4
Epidemiology	78.1	100.0	87.7	78.6	100.0	88.0
Event	100.0	100.0	100.0	91.9	100.0	95.8
Goods	70.4	76.0	73.1	97.1	100.0	98.5
Character	100.0	100.0	100.0	83.3	83.3	83.3
Overall	74.6	91.5	82.2	85.6	95.0	90.0

[a]Not applicable.

Knowledge Access, Sustainability, and Possible Applications

Naming Convention

For considerations of readability and interoperability, we followed the RDF naming convention, which helps to quickly locate and understand the topic and the meaning of each triple. The convention is composed of three major parts.

The first is the resource identifier, in which each resource (ie, concept, entity, property) is identified by a global ID that is an integer number prefixed by a letter. That is, classes are prefixed by C (eg, C1), entities are prefixed by R (eg, R122), and properties are prefixed by P (eg, P31). The second is the uniform resource identifier (URI) pattern. All URIs should follow a pattern such as [URL]/[graphname]/[type]/[resource], where graphname is the name of the subgraph (eg, medical, research), type takes on an enumerable value representing the URI type (ie, class, resource, property), and resource is the global identifier described in the resource identifier part. The third part is the predicate usage; the COVID-19 KGs use the set of predicates shown in Table 5 to illustrate the schema model.

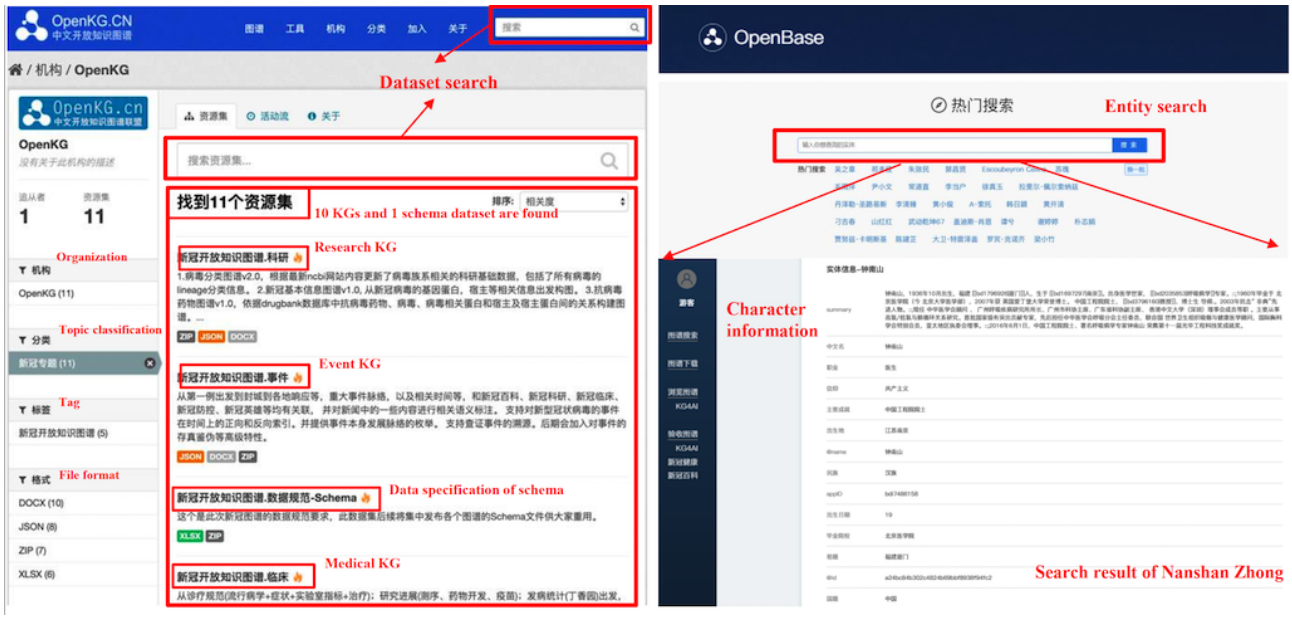
Table 5. Primary predicates used in the OpenKG-COVID19 schemata.

Predicate	Description
rdf:label	Local name statement of all URLs
rdfs:subClassOf	The hypernym-hypernym relationship between two classes
rdfs:domain	Domain class of a property
rdfs:range	Range class or literal data types of a property, which can be multivalued
owl:sameA	Synonym relationship between two resources

Search and Browse Interfaces

Since the published KGs in OpenKG-COVID19 comply with the creative commons-by share-alike license, users can feel free to download any of them [47]. The left part of Figure 6 shows a snapshot of the data set search interface, where 10 KGs and a schema data set about OpenKG-COVID19 are found. Moreover, users can search for a particular entity and browse the detailed information of that entity in the OpenBase website [48]. As shown in the right part of Figure 6, the search results contain various properties of Nanshan Zhong, a famous doctor combating the COVID-19 epidemic in China.

Figure 6. Data set search interface (left) and entity search interface (right).



SPARQL Endpoint

The SPARQL endpoint [49] of OpenKG-COVID19 is built upon a scalable graph database, gStore [50], which provides extendable distributed storage management as well as efficient implementations of complex queries and update operations based on SPARQL for RDF data sets with up to billions of triples. Users can submit SPARQL queries to the endpoint where relevant results are returned in the form of a table. Users can also choose to download the results packaged in a JavaScript Object Notation (JSON) file by clicking “Click to Download.” As of May 22, 2020, we have recorded over 20,000 accesses to the endpoint.

Sustainability and Knowledge Review

OpenKG-COVID19 KGs are maintained by the OpenKG community. We are collecting questionnaires considering users’ needs and updating our KGs accordingly. COVID-19 KGs are particularly important for timely updates because users’ needs may change as the epidemic develops (eg, from source to treatment).

The data quality as well as the interlinking quality of OpenKG-COVID19 are manually evaluated. OpenBase is a knowledge crowdsourcing platform powered by blockchain technologies for provenance tracking and credit incentive. We uploaded a part of the data that may contain errors due to the sampling method, and created many microtasks for reviewing the correctness of triples. The reviewers were volunteers certified by possession of one specific domain knowledge. They were able to not only review the KG data but to also commit data corrections. All volunteers participating in knowledge reviewing via either a web-based interface or the WeChat mini app (see Figure 7) received a corresponding reward of credit for their contributions to improving our KGs.

Figure 7. Knowledge review on the web (left) and WeChat mini app (right) of OpenBase.

Possible Intelligent Applications

OpenKG-COVID19 is the basis of various intelligent applications, whose release will help to fight against this global plague. OpenKG-COVID19 benefits from intelligent question answering, semantic search, recommendation systems, as well as the abilities for visualization, mining more associations, predicting future events, and assisting in decision-making. More specifically, we here take the event KG, research KG, medical KG, and overall OpenKG-COVID19 as examples.

The event KG includes the forward and reverse indexing of events about COVID-19 in time, and provides the development context of a series of events, which can support the verification and traceability of hot events. Furthermore, the event KG combined with blockchain technology could identify whether or not an event is true.

Based on the research KG, Huawei Cloud has developed a personalized visual query system, displaying knowledge points and their relations, which can quickly trace the source of information and directly locate relevant documents and paragraphs. The research KG facilitates scientific research on virus mechanisms and viral protein interactions, and assists drug developers in more accurate and effective drug target research and vaccine development.

Starting from the cases of diagnosis and treatment to research progress, the medical KG is developed by extracting knowledge from the existing standard documents and the web. The epidemiology, symptoms, laboratory indicators, treatments, drug development, and vaccines of COVID-19 could be conveniently consulted making use of question answering based on the medical KG. Drugs that alleviate symptoms and potential therapeutic drugs, such as the repurposing of old drugs for a new use, can also be mined by the medical KG.

Moreover, OpenKG-COVID19 is an enabler to accelerate the development of bioinformatics. The network structures of COVID-19 KGs can be used to predict relations such as host-virus, drug-virus, or interactions between viruses and the host protein, which will help to reveal the underlying mechanism of COVID-19. In particular, the combination of protein-protein interactions, drug–protein target interactions, and the polypharmacy side effects could predict unknown side effects.

Discussion

Principal Results

In this study, we constructed OpenKG-COVID19, one of the largest existing KGs about COVID-19. We first presented the schema design process of OpenKG-COVID19. We then introduced the comprehensive techniques for knowledge extraction and knowledge fusion. Moreover, we provided an evaluation of the quality of OpenKG-COVID19. This paper also provides an introduction of various access interfaces covering searching, browsing, querying, and knowledge review, and discusses the possible applications of OpenKG-COVID19. Our efforts can benefit KG, biomedicine, and many other communities. New knowledge for the 10 KGs will be updated continuously through the processes described above to maintain and update OpenKG-COVID19 for improving its quality and coverage.

Limitations

Although OpenKG-COVID19 is updated continuously, the update frequency is not daily, which may result in some information not being up to date, causing inconvenience for downstream tasks. Moreover, it is also very necessary to control the data set version, which is future work to be considered.

We randomized a chosen sample of facts for each property defined in OpenKG-COVID19 to evaluate the data quality. In some cases, the number of samples may be small, which will lead to a less reliable evaluation result. Therefore, we plan to further improve the quality of data by selecting a new method to sample more triples of each property.

Conclusion

A KG is an effective technique to provide well-organized data, and is also beneficial for intelligent question answering, semantic search, recommendation system, visualization analysis, and decision-making support. OpenKG-COVID19 includes rich and diverse topics of COVID-19, covering 10 aspects ranging from encyclopedia, concept, medical, health, prevention, goods, research, epidemiology, and character to events. The publishing and maintenance of OpenKG-COVID19 can help researchers around the world to better understand, study, and even fight COVID-19.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (grant 22120220069) and the National Nature Science Foundation of China (grant 62176185).

Authors' Contributions

WH guided the project, advised on the construction of each subgraph, and wrote the paper. HD participated in construction of OpenKG-COVID19, performed the experiments, and wrote the paper. GQ, HC, and WH participated in the construction of OpenKG-COVID19 and gave advice during the project. All authors reviewed the manuscript and approved the final version.

Conflicts of Interest

None declared.

References

1. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. Severe acute respiratory syndrome-related coronavirus?the species and its viruses, a statement of the coronavirus study group. BioRxiv. 2020 Feb 11. URL: https://www.biorxiv.org/content/10.1101/2020.02.07.937862v1 [accessed 2022-04-29]
2. Yang Z, Zeng Z, Wang K, Wong S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020 Mar;12(3):165-174. [doi: 10.21037/jtd.2020.02.64] [Medline: 32274081]
3. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020 Nov 11;10(1):19549. [doi: 10.1038/s41598-020-76550-z] [Medline: 33177550]
4. Mohamed SK, Nováček V, Nounu A. Discovering protein drug targets using knowledge graph embeddings. Bioinformatics 2020 Jan 15;36(2):603-610. [doi: 10.1093/bioinformatics/btz600] [Medline: 31368482]
5. Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M. COBERT: COVID-19 question answering system using BERT. Arab J Sci Eng 2021 Jun 23:1-11 [FREE Full text] [doi: 10.1007/s13369-021-05810-5] [Medline: 34178569]
6. Zhang Y, Zhang X, Hu Y, Wang G, Yan R. Wulai-qa: Web understanding and learning with ai towards document-based question answering against covid-19. 2021 Presented at: 14th ACM International Conference on Web Search and Data Mining; March 8-12, 2021; Jerusalem, Israel p. 898-901. [doi: 10.1145/3437963.3441707]
7. Ding K, Han H, Li L, Menglin Y. Research on question answering system for covid-19 based on knowledge graph. 2021 Presented at: 2021 40th Chinese Control Conference (CCC); July 26-28, 2021; Shanghai, China p. 4659-4664. [doi: 10.23919/ccc52363.2021.9550437]
8. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. Dbpedia: A nucleus for a web of open data. In: The Semantic Web. ISWC ASWC 2007 2007. Lecture Notes in Computer Science, vol 4825. Berlin, Heidelberg: Springer; 2007:735.
9. Suchanek F, Kasneci G, Weikum G. Yago: a core of semantic knowledge. 2007 Presented at: WWW '07: Proceedings of the 16th international conference on World Wide Web; May 8-12, 2007; Banff, Alberta p. 697-706. [doi: 10.1145/1242572.1242667]
10. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. 2008 Presented at: 2008 ACM SIGMOD international conference on Management of Data; June 10-12, 2008; Vancouver, BC p. 1247-1250. [doi: 10.1145/1376616.1376746]
11. Vrandečić D, Krötzsch M. Wikidata. Commun ACM 2014 Sep 23;57(10):78-85. [doi: 10.1145/2629489]
12. Niu X, Sun X, Wang H, Rong S, Qi G. Zhishi.me - weaving Chinese linking open data. In: The Semantic Web – ISWC 2011. Lecture Notes in Computer Science, vol 7032. Berlin, Heidelberg: Springer; 2011:205-220.
13. Miller G. In: Fellbaum C, editor. WordNet: An electronic lexical database. Cambridge, MA: MIT press; 1998.
14. Navigli R, Ponzetto S. Babelnet: Building a very large multilingual semantic network. 2010 Presented at: 48th Annual Meeting of the Association for Computational Linguistics; July 2010; Uppsala, Sweden p. 216-225.
15. Chiarcos C, Declerck T, Mccrae J. Linguistic linked open data (LLOD). Introduction and overview. 2013 Presented at: 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data; September 2013; Pisa, Italy. [doi: 10.1007/978-3-030-30225-2_1]
16. Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, et al. Linked open drug data for pharmaceutical research and development. J Cheminform 2011 May 16;3(1):19. [doi: 10.1186/1758-2946-3-19] [Medline: 21575203]
17. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, et al. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 2012 Nov;17(21-22):1188-1198 [FREE Full text] [doi: 10.1016/j.drudis.2012.05.016] [Medline: 22683805]
18. Callahan A, Cruz-Toledo J, Ansell P, Dumontier M. Bio2rdf release 2: improved coverage, interoperability and provenance of life science linked data. In: Cimiano P, Corcho O, Presutti V, Hollink L, Rudolph S, editors. The Semantic Web: Semantics and Big Data. ESWC 2013. Lecture Notes in Computer Science, vol 7882. Berlin, Heidelberg: Springer; 2013:14.

19. Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, et al. CORD-19: The Covid-19 Open Research Dataset. ArXiv. 2020 Apr 22. URL: https://arxiv.org/abs/2004.10706 [accessed 2022-04-29]
20. Colavizza G, Costas R, Traag VA, van Eck NJ, van Leeuwen T, Waltman L. A scientometric overview of CORD-19. PLoS One 2021;16(1):e0244839 [FREE Full text] [doi: 10.1371/journal.pone.0244839] [Medline: 33411846]
21. Domingo-Fernández D, Baksi S, Schultz B, Gadiya Y, Karki R, Raschka T, et al. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics 2021 Jun 09;37(9):1332-1334 [FREE Full text] [doi: 10.1093/bioinformatics/btaa834] [Medline: 32976572]
22. Zhang P, Bu Y, Jiang P, Shi X, Lun B, Chen C, et al. Toward a coronavirus knowledge graph. Genes 2021 Jun 29;12(7):998 [FREE Full text] [doi: 10.3390/genes12070998] [Medline: 34209818]
23. COVIDGRAPH – a COVID-19 knowledge graph. HealthECCO. URL: https://covidgraph.org/#applications [accessed 2022-04-23]
24. Payne S, Large S, Jarrett N, Turner P. Written information given to patients and families by palliative care units: a national survey. Lancet 2000 May 20;355(9217):1792. [doi: 10.1016/S0140-6736(00)02272-8] [Medline: 10832835]
25. Gong F, Chen Y, Wang H, Lu H. On building a diabetes centric knowledge base via mining the web. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):49 [FREE Full text] [doi: 10.1186/s12911-019-0771-6] [Medline: 30961582]
26. Swanson-Kauffman K. There should have been two: nursing care of parents experiencing the perinatal death of a twin. J Perinat Neonatal Nurs 1988 Oct;2(2):78-86. [doi: 10.1097/00005237-198810000-00010] [Medline: 3418501]
27. OpenKG. URL: http://openkg.cn [accessed 2022-02-13]
28. COVID-19 schema. OpenKG. URL: http://openkg.cn/dataset/covid-19-schema [accessed 2022-04-22]
29. BabelNet. URL: https://babelnet.org/ [accessed 2022-02-13]
30. Wu T, Wang H, Qi G, Zhu J, Ruan T. On building and publishing Linked Open Schema from social web sites. J Web Semant 2018 Aug;51:39-50. [doi: 10.1016/j.websem.2018.05.002]
31. National Center for Biotechnology Information. URL: https://www.ncbi.nlm.nih.gov [accessed 2022-02-13]
32. GISAID. URL: https://www.gisaid.org/ [accessed 2022-02-13]
33. Database Commons: A Catalog of Biological Databases. URL: https://ngdc.cncb.ac.cn/databasecommons/database/id/1796 [accessed 2022-02-13]
34. Drug-Bank. Building the foundation for better health outcomes. URL: https://go.drugbank.com/ [accessed 2022-02-13]
35. Nextstrain. URL: https://nextstrain.org/ [accessed 2022-02-13]
36. Leskovec J. Stanford Biomedical Network Dataset Collection. Stanford University. URL: http://snap.stanford.edu/biodata/ [accessed 2022-02-13]
37. viPR Virus Pathogen Resource. URL: https://www.viprbrc.org/ [accessed 2022-02-13]
38. VirHostNet 3.0: towards systems biology of virus/host interactions. URL: http://virhostnet.prabi.fr/ [accessed 2022-02-13]
39. VISDB Viral Integration Site DataBase. URL: https://bioinfo.uth.edu/VISDB/ [accessed 2022-02-13]
40. Li X, Feng J, Meng Y, Han Q, Wu F, Li J. A unified MRC framework for named entity recognition. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 2020; online p. 5849-5859. [doi: 10.18653/v1/2020.acl-main.519]
41. 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: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy p. 1340-1350. [doi: 10.18653/v1/p19-1129]
42. Kenton J, Toutanova L. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 Presented at: 17th Annual 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]
43. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. 2009 Presented at: Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP; August 2009; Suntec, Singapore p. 1003-1011. [doi: 10.3115/1690219.1690287]
44. Hu W, Qu Y. Falcon-AO: A practical ontology matching system. J Web Semant 2008 Sep;6(3):237-239. [doi: 10.1016/j.websem.2008.02.006]
45. Chai C, Li G, Li J, Deng D, Feng J. A partial-order-based framework for cost-effective crowdsourced entity resolution. VLDB J 2018 Jun 12;27(6):745-770. [doi: 10.1007/s00778-018-0509-6]
46. Tao Y. Entity matching with active monotone classification. 2018 Presented at: 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems; June 10-15, 2018; Houston, TX p. 49-62. [doi: 10.1145/3196959.3196984]
47. Coronavirus dataset. OpenKG. URL: http://openkg.cn/dataset?groups=coronavirus [accessed 2022-04-22]
48. OpenBase. URL: http://openbase.openkg.cn/ [accessed 2022-02-13]
49. Query endpoint. OpenKG. URL: http://pkubase.gstore.cn [accessed 2022-02-13]
50. Zou L, Özsu MT, Chen L, Shen X, Huang R, Zhao D. gStore: a graph-based SPARQL query engine. VLDB J 2013 Sep 28;23(4):565-590. [doi: 10.1007/s00778-013-0337-7]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformer
JSON: JavaScript Object Notation
KG: Knowledge Graph
NCBI: National Center for Biotechnology Information
RDF: Resource Description Framework
URI: uniform resource identifier

Edited by T Hao; submitted 07.03.22; peer-reviewed by S He, Z Huang; comments to author 20.04.22; revised version received 23.04.22; accepted 26.04.22; published 13.05.22.

Please cite as:
Wang H, Du H, Qi G, Chen H, Hu W, Chen Z
Construction of a Linked Data Set of COVID-19 Knowledge Graphs: Development and Applications
JMIR Med Inform 2022;10(5):e37215
URL: https://medinform.jmir.org/2022/5/e37215
doi:10.2196/37215
PMID:35476822

Original Paper

An Analysis of French-Language Tweets About COVID-19 Vaccines: Supervised Learning Approach

Romy Sauvayre[1,2], PhD, HDR; Jessica Vernier[1], MA; Cédric Chauvière[2,3], PhD

[1]Laboratoire de Psychologie Sociale et Cognitive, Université Clermont Auvergne, Centre national de la recherche scientifique, Clermont-Ferrand, France

[2]Polytech Clermont, Clermont Auvergne INP, Aubiere, France

[3]Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, Centre national de la recherche scientifique, Clermont-Ferrand, France

Corresponding Author:
Romy Sauvayre, PhD, HDR
Polytech Clermont
Clermont Auvergne INP
2, avenue Blaise Pascal
Aubiere, 63178
France
Phone: 33 4 73 40 55 4
Fax: 33 4 73 40 75 1
Email: romy.sauvayre@uca.fr

Abstract

Background: As the COVID-19 pandemic progressed, disinformation, fake news, and conspiracy theories spread through many parts of society. However, the disinformation spreading through social media is, according to the literature, one of the causes of increased COVID-19 vaccine hesitancy. In this context, the analysis of social media posts is particularly important, but the large amount of data exchanged on social media platforms requires specific methods. This is why machine learning and natural language processing models are increasingly applied to social media data.

Objective: The aim of this study is to examine the capability of the CamemBERT French-language model to faithfully predict the elaborated categories, with the knowledge that tweets about vaccination are often ambiguous, sarcastic, or irrelevant to the studied topic.

Methods: A total of 901,908 unique French-language tweets related to vaccination published between July 12, 2021, and August 11, 2021, were extracted using Twitter's application programming interface (version 2; Twitter Inc). Approximately 2000 randomly selected tweets were labeled with 2 types of categorizations: (1) arguments for (pros) or against (cons) vaccination (health measures included) and (2) type of content (scientific, political, social, or vaccination status). The CamemBERT model was fine-tuned and tested for the classification of French-language tweets. The model's performance was assessed by computing the F1-score, and confusion matrices were obtained.

Results: The accuracy of the applied machine learning reached up to 70.6% for the first classification (pro and con tweets) and up to 90% for the second classification (scientific and political tweets). Furthermore, a tweet was 1.86 times more likely to be incorrectly classified by the model if it contained fewer than 170 characters (odds ratio 1.86; 95% CI 1.20-2.86).

Conclusions: The accuracy of the model is affected by the classification chosen and the topic of the message examined. When the vaccine debate is jostled by contested political decisions, tweet content becomes so heterogeneous that the accuracy of the model drops for less differentiated classes. However, our tests showed that it is possible to improve the accuracy by selecting tweets using a new method based on tweet length.

(JMIR Med Inform 2022;10(5):e37831) doi:10.2196/37831

KEYWORDS

social media; natural language processing; public health; vaccine; machine learning; CamemBERT language model; method; epistemology; COVID-19; disinformation; language model

Introduction

Background

The COVID-19 pandemic has profoundly affected our society and social activity worldwide. Part of this change is perceptible through messages exchanged on social media platforms, specifically on the topic of vaccination. Since the measles, mumps, and rubella vaccine controversy in 1998 [1], vaccine hesitancy has grown on the internet [2,3] and subsequently on social media platforms such as Facebook and Twitter [4,5]. In the same way, as the pandemic progressed, disinformation, "fake news," and conspiracy theories spread [6] through many parts of society. However, the disinformation spreading through social media is, according to the literature, "potentially dangerous" [7] and is one of the causes of increased COVID-19 vaccine hesitancy [8,9]. Another cause mentioned in the literature is the loss of confidence in science among the public [10].

In this context, social media analysis is particularly important, but the large amount of data exchanged over social networks requires specific methods. This is why machine learning and natural language processing (NLP) models are becoming increasingly popular for studying social media data. The most used and "most promising method" [11] is sentiment analysis. For example, sentiment analyses were conducted on messages posted on Twitter (tweets) to measure the opinions of Americans regarding vaccines [12] and evaluate the rate of hate tweets among Arab people [13]. Additionally, another method, opinion mining, is used and has obtained an equal level of maturity [14]. Both methods attempt to identify and categorize subjective content in text, but it is not an easy task to correctly identify such concepts (opinion, rumor, idea, claim, argument, emotion, sentiment, and affect). The fields of psychology and philosophy have extensively studied these concepts but have raised the difficulty of defining their boundaries. This is why stance detection has grown to be considered "a subproblem of sentiment analysis" [15]. In addition, according to Visweswaran et al [16], performing a sentiment analysis on tweets is a challenge because tweets contain short text (280 characters or less), abbreviations, and slang terms. However, few studies focus on the difficulties encountered by a neural network according to the chosen categories [17]. The aim of this paper is to provide additional methodological reflection.

Objective

The aim of this study is to examine the capability of the CamemBERT model to faithfully predict the elaborated categories while considering that tweets about vaccination are often ambiguous, sarcastic, or irrelevant to the studied topic. Based on the resulting analysis, this paper aims to provide a methodological and epistemological reflection on the analysis of French-language tweets related to vaccination.

A State-of-the-art French-Language Model

The CamemBERT model was released in 2020 and is considered one of the state-of-the-art French-language models [18] (together with its close "cousin" flauBERT [19]). It makes use of the Robustly Optimized BERT Pretraining Approach architecture of Liu et al [20], which is an improved variant of the famous Bidirectional Encoder Representations From Transformers (BERT) architecture of Devlin et al [21]. The BERT family of models consists of general, multipurpose, pretrained models that may be used for different NLP tasks, including the following: classification, question answering, and translation. They rely heavily upon transformers, which have radically changed the performance of NLP tasks since their introduction by Google researchers in 2017 [22]. They have been pretrained on a large corpus ranging from gigabits to terabits of data, using considerable computing resources.

Although multilingual models are plentiful, they usually lag behind their monolingual counterparts. This is why, in this study, we chose to employ a monolingual model to classify French-language tweets. As far as we are concerned, CamemBERT comes in 6 different "flavors," ranging from small models with 110 million parameters trained on 4 GB of text up to mid-size models with 335 million parameters trained on 135 GB of text. After testing them, we found that better results were obtained with the largest size model that was pretrained on the Criss-Cross Network corpus.

All these models require fine-tuning on specific data to achieve their full potential. Fine-tuning or transfer learning have been common and successful practices in computer vision for a long time, but it is only in the last 3 years or so that the same approaches have become effective for solving NLP problems on specific data. This approach can be summarized in the following 3 steps:

1. A model language such as BERT is built in an unsupervised manner using a large database, removing the need to label data.
2. A specific head (such as dense neural network layers) is added to the previous model to make it task-specific.
3. The new model is trained in its entirety with a small learning rate on specific data.

The first step is usually performed by large companies, such as Google or Facebook, or public research centers that make their model freely available on internet platforms. The second and third steps form a process that is generally referred to as *fine-tuning*, and this is what we will do in this study.

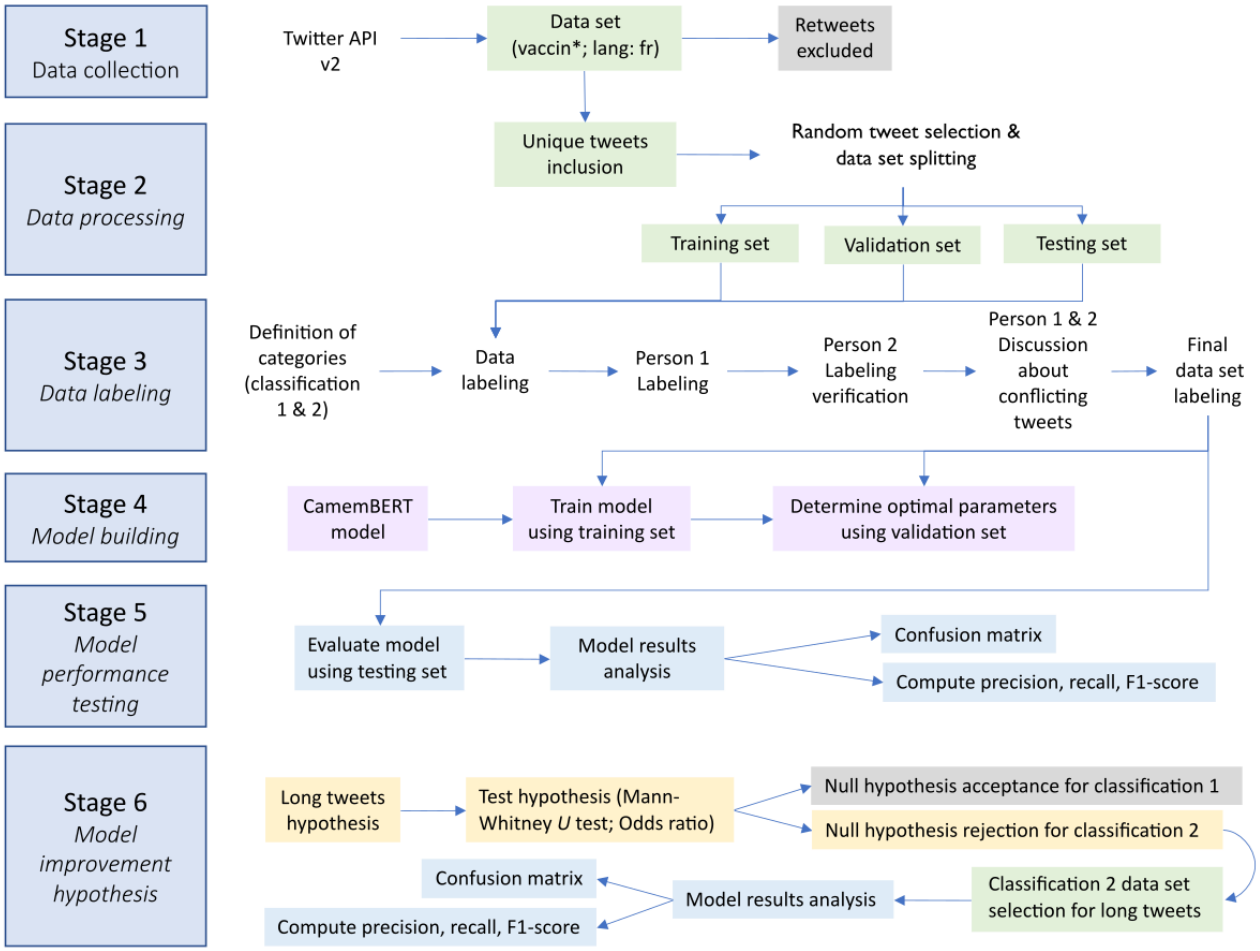
Methods

Data Collection

French-language tweets published between July 12, 2021, and August 11, 2021, were extracted using the Twitter application programming interface ([API] version 2; Twitter Inc; Figure 1) with a Python (Python Software Foundation) script request (vaccin lang: fr), and several elements (tweet content, tweet ID, author ID, and creation date) were stored in a document-oriented database (MongoDB, MongoDB Inc). As queries can only contain a limited number of terms (1024 characters), it was more relevant to search for the word *vaccin* ("vaccine"), knowing that related terms were included by the Twitter API version 2 search tools since November 15, 2021, rather than selecting a nonexhaustive keyword list. Indeed, Twitter's query tool collected all words containing the base word *vaccin* in

French (ie, *vaccin*, *vaccins*, *vaccination*, *vaccinations*, *vaccinat*, *vacciner*, *vaccinés*, *vaccinées*, *vaccinerait*, *vaccineraient*, *pro-vaccin*, *anti-vaccin*, *#vaccin*, *#vaccinationobligatoire*). The goal of this approach was to collect all tweets containing the base word *vaccin* to explore their content using a bottom-up approach without additional inclusion or exclusion criteria. A total of 1,782,176 tweets were obtained, including 901,908 unique tweets (29,094 tweets per day) published by 231,373 unique users. To fully test the CamemBERT model, only unique tweets were included in the analysis. When dealing with the analysis of text (such as tweets), it is important to keep a large amount of variability (eg, vocabulary, syntax, and length) to strengthen deep learning algorithms. This variability will guarantee the power of model generalization. This is why, in this study, the 1851 tweets that comprise the data set were drawn randomly from a set of 901,908 unique tweets.

Figure 1. Flow chart of methodology steps. API v2: application programming interface version 2.



Labeling

A total of 1851 unique tweets were randomly selected and manually labeled by 2 people (1451 for training and validation and 400 for testing). When doubt arose about labeling, which occurred for 87 of the 1851 tweets (4.7%), a discussion occurred to determine the relevant label for each tweet (see examples in Multimedia Appendix 1). Note that no duplicates were identified by the automated verification performed.

A total of 2 classifications were developed to examine arguments for (pros) or against (cons) vaccination (health measures included) and examine the type of tweet content (scientific, political, social, or vaccination status). The classifications and definitions used to label tweets are provided in Table 1 with translated examples of tweets for each label. In accordance with Twitter's terms of use under the European General Data Protection Regulation, original tweets cannot be shared [23]. Therefore, the translations have been adjusted to ensure the anonymity of Twitter users.

Table 1. Classification criteria for tweets and definitions.

Type of tweet	Definition	Translated examples (French to English)
Classification problem 1		
Unclassifiable	Unclassifiable or irrelevant to the topics of vaccination or health measures	The Emmanuel Macron effect
Noncommittal	Neutral or without explicit opinion on vaccination and/or the health pass	I have to ask my doctor for the vaccine
Pros	Arguments in favor of the health pass Arguments in favor of the COVID-19 vaccine and/or the health pass (efficiency, safety, relevance)	Personally, I am vaccinated so nothing to fear, on the other hand, good luck to all the anti-vaccine, you will not have the choice now??
Cons	Arguments against vaccination or doubts about the effectiveness of COVID-19 vaccines, fear of side effects, and refusal to obtain the health pass	I am against the vaccine I am not afraid of the virus but I am afraid of the vaccine
Classification problem 2		
Unclassifiable	Irrelevant to the topic or unclassifiable	A vaccine
Scientific	Scientific or pseudoscientific content that uses true beliefs or false information	The vaccine is 95% efficient, a little less in fragile people. The risk is not zero, but a vaccinated person has much less chance of transmitting the virus.
Political	Comments on legal or political decisions about vaccination or health measures	Basically the vaccine is mandatory, shameful LMAO
Social	Comments, debates, or opinions on the report to other members of society	“Pro vaccine” you have to also understand that there are people who do not want to be vaccinated.
Vaccination status	Explicit tweet about the vaccination status of the tweet author Comments on the symptoms experienced after COVID-19 vaccination Explicit refusal to receive a COVID-19 vaccine	Example 1: I am very glad to have already done my 2 doses of the vaccine, fudge Example 2: I don't want to get vaccinated. Why? Well, you know, we don't know what's in this vaccine, it can be dangerous.

Classification Method

This study followed the general methodology of machine learning to guarantee a rigorous building of the model. To ensure that the model did not overfit or underfit the data set, the following steps were taken:

1. The data set was divided into training (n=1306), validation (n=145), and testing (n=400) data sets.
2. The training loss was represented as a function of the number of epochs to monitor the correct learning of the model and select its optimal value.
3. The validation accuracy is represented as a function of the number of epochs to ensure that the model was not overfitting or underfitting the data.
4. The final model was evaluated on a testing data set that had not been previously used to build or validate the model.

A total of 2 fully connected dense neural network layers with 1024 and 4 neurons (for classification problem 1) or 5 neurons (for classification problem 2) were added to the head of the CamemBERT model, adding another 1.6 million parameters. Furthermore, to prevent overfitting, a 10% dropout was applied between those 2 layers. A small learning rate of 2×10^{-5} was used for fine-tuning, and adaptive moment estimation with a decoupled weight decay regularization [24] was chosen as the optimizer (see full code used on GitHub [25]). The parameters were adjusted by minimizing the cross-entropy loss, which is a common choice when dealing with a classification problem.

Fine-tuning was performed on a data set consisting of the 1451 labeled French-language tweets, 90% (n=1306) of which were used for training and the remaining 10% (n=145) for validation. Once the model was built, it was tested on a new set of 400 labeled tweets from which a statistical analysis was performed. A total of 2 classification models were built from the same data set, 1 with 4 labels (unclassifiable, neutral, positive, or negative) related to a tweet author’s opinion about vaccination and 1 with 5 labels related to the type of content in a tweet (unclassifiable, scientific, political, social, vaccination status, or symptoms). The proportion of tweets classified into each label for these 2 problems is given in Table 2. We see that the data set is slightly imbalanced. As such, it does not require special treatment.

One of the main hyperparameters to be tuned for the training of the model is the number of epochs. As a rule of thumb, to prevent overfitting, the number of epochs is usually chosen based on when the abruptness of the slope of the loss changes while maintaining a low rate of misclassification on the validation data set. Figure 2 shows that 7 epochs should lead to the best result.

This was confirmed by computing the precision, recall, and F1-score at 3 different epochs (7, 15, and 20), as shown in Table 3. The reported results were computed on the test data set with 400 tweets. The average results over the classes were weighted to account for imbalanced classes in the data set. As expected, the highest score was obtained with 7 epochs, however, not by a wide margin (Table 3).

A similar study for the second classification problem determined that 6 epochs were enough to prevent overfitting. The performance of the model was also measured by computing the weighted precision, recall, and F1-score, as shown in Table 4.

The size of the data set is quite similar to those of Kummervold et al [17] (1633 tweets for training and 544 for testing) and Benítez-Andrades et al [26] (n=1400 for training and n=600 for testing). Furthermore, the benefit of using a pretrained model such as the CamemBERT is that a large data set is not required to obtain good results. We also tried to build a neural network model from scratch with the same data set, but the classification performance of the model was significantly lower than the results presented in this paper with the CamemBERT model. For classification problem 1, we reached an accuracy of 33% (versus 59% with the pretrained model) and for classification problem 2, we reached an accuracy of 40% (versus 67.6% with the pretrained model).

Table 2. The proportion of tweets assigned to each label in the data set for classification problems 1 and 2 (n=1451).

Classification problem	Tweets
Classification problem 1, n (%)	
Unclassifiable	189 (13)
Neutral	354 (24.4)
Positive	392 (27)
Negative	516 (35.6)
Classification problem 2, n (%)	
Unclassifiable	226 (15.6)
Scientific	441 (30.4)
Political	316 (21.8)
Social	353 (24.3)
Vaccination status	115 (7.9)

Figure 2. Training loss (a) and validation accuracy (b) of the model over 20 epochs for classification problem 1.

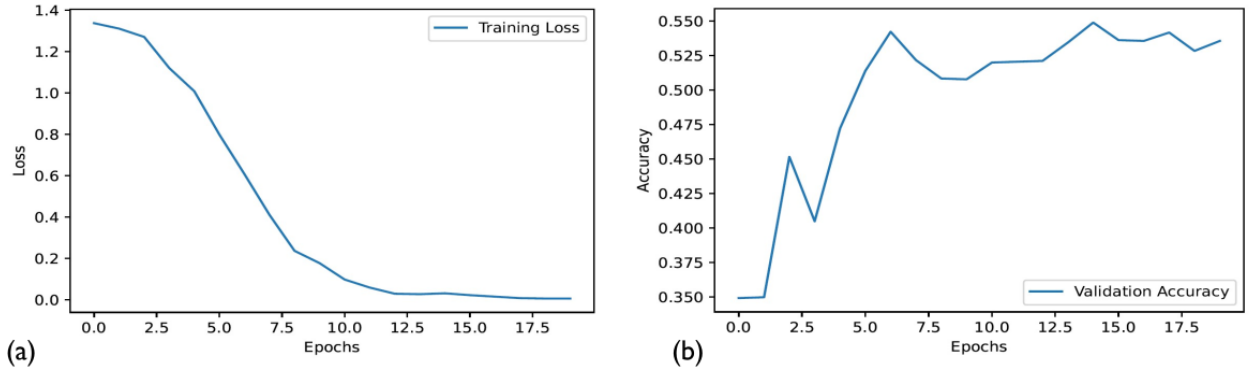


Table 3. Classification performance of the model for classification problem 1.

Epochs, n	Precision[a]	Recall[a]	F1-score[a]
7	59	55.3	55.3
15	56.6	53	53.2
20	56.9	54.5	55.2

[a]These data are provided as percentages.

Table 4. Classification performance of the model for classification problem 2.

Epochs, n	Precision[a]	Recall[a]	F1-score[a]
6	67.6	64.5	62.9
15	62.7	62.8	61.3
20	60.6	59.5	56.5

[a]These data are provided as percentages.

Results

Statistical Analysis

From the results of the previous section, we see that it is significantly more difficult to build a performant classifier based on the 4 vaccine sentiment labels (unclassifiable, noncommittal, pros, and cons), with the maximum F1-score reaching 55.3% in this case. On the other hand, the classifier built from the same tweets but with 5 different labels based on content type (unclassifiable, scientific, political, social, vaccination status, or symptoms) achieved a much higher F1-score (62.9%).

To analyze the strength and weakness of a model more specifically, it is always instructive to represent it using a confusion matrix [27], as shown in Figure 3.

Since the values in these matrices are percentages, their interpretation requires some care. For the first problem, summing figures line-by-line in the matrix shows that out of 100 tweets from the test data set, on average, 11.25 are unclassifiable, 35.50 are noncommittal, 13.25 are pros, and 40.00 are cons. It is then possible to compute the proportion of tweets correctly classified by the model, label-by-label. The results are shown in Table 5. We see that the model can accurately classify the tweets labelled as pros and cons. It misclassifies a large number of the unclassifiable tweets and, to a lesser extent, noncommittal tweets. Looking back to the confusion matrix, for the last 2 labels, we observe that the model tends to classify the tweets as being pros.

For the second problem, as expected, in line with the higher F1-score found in the previous section, the model achieves much better classification performance. It excels at classifying scientific and political tweets and is also good at classifying social tweets. It still has some difficulties classifying unclassifiable tweets and, in a larger proportion, vaccination status tweets. Looking back to the confusion matrix, for the last 2 labels, we observe that the model tends to classify them as being social tweets.

Figure 3. Confusion matrix for classification problems 1 and 2 (n=400).

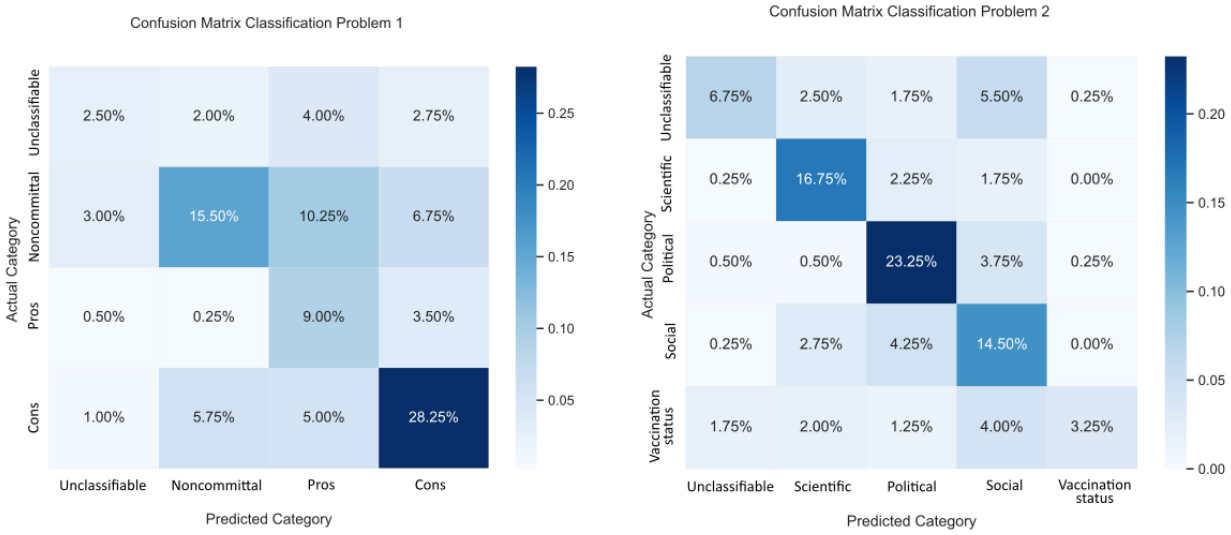


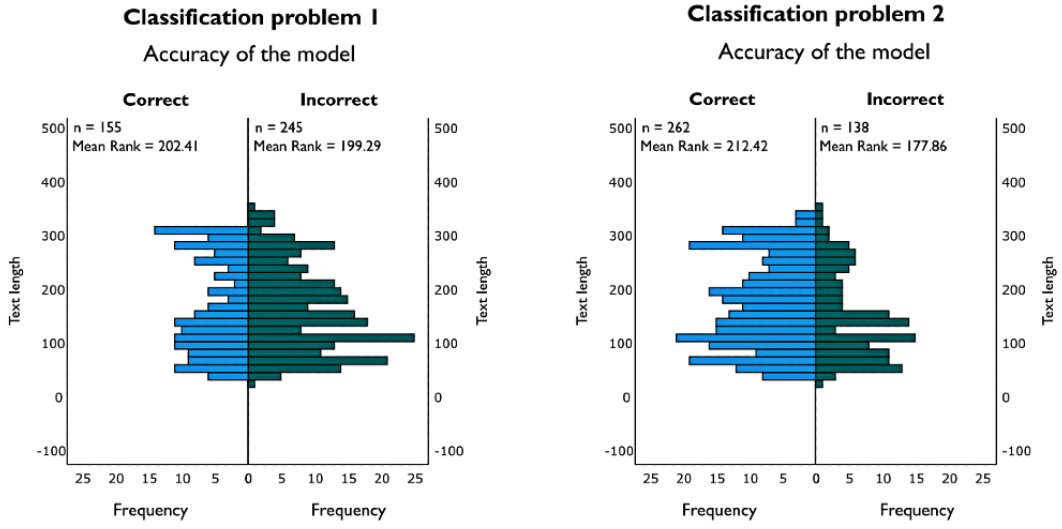
Table 5. The number of tweets correctly classified for each label in classification problems 1 and 2 (n=400).

Classification problem	Tweets
Classification problem 1, n (%)	
Unclassifiable	10 (22.2)
Noncommittal	62 (43.7)
Pros	36 (67.9)
Cons	113 (70.6)
Classification problem 2, n (%)	
Unclassifiable	27 (40.3)
Scientific	67 (79.8)
Political	93 (82.3)
Social	58 (66.7)
Vaccination status	13 (26.5)

Text Size Analysis

To improve the performance of the fine-tuned CamemBERT model, a hypothesis about the influence of tweet length on model accuracy was tested. A Mann-Whitney *U* test generated statistically significant results for classification problem 2 (*U*=21,202; *P*=.004) but not for classification problem 1 (*U*=19,284; *P*=.79). As Figure 4 shows, the correctly predicted tweets are significantly longer for classification problem 2. A second analysis carried out on a dichotomous variable created from the tweet text length (greater than or less than 170 characters) confirmed this significance for classification problem 2. A tweet was 1.86 times more likely to be incorrectly predicted by the model if it contained less than 170 characters (odds ratio [OR] 1.86; 95% CI 1.20-2.86). Therefore, the significance obtained using these 2 analyses (Mann-Whitney *U* test and OR) allows us to rigorously validate [28] our hypothesis.

Figure 4. Tweet text length as a function of the accuracy of the fine-tuned CamemBERT model conducted on classification problems 1 and 2 (Mann-Whitney *U* test).



Long Tweet Test

The finding of the previous section is further supported after carrying out the following experiment. Tweets with more than 170 characters were selected from the 400-tweet data set. Classification model 2 was then tested with these 168 tweets to see if its accuracy increased.

As shown in Table 6, the accuracy improved from 64.5% to 73.2% (an 8.7% increase), confirming our hypothesis. The F1-score also increased by approximately the same amount.

The confusion matrix generated from the comparison between the model-classified and the manually classified 168 long tweets is shown in Figure 5. From this matrix, it is possible to compute the percentage of correct classifications for each label, the results of which are shown in Table 7. The increase in accuracy is significant for the vaccination status label (an increase of 9.2%), followed by the political label (an increase of 7.7%) and the unclassifiable label (an increase of 6%).

As already pointed out using the Mann-Whitney *U* test and OR, the model for the second problem has much better classification performance with long tweets. It should be noted that the rate of correct classification of political tweets reached an impressive 90% (45/50).

Table 6. Classification performance of the model for classification problem 2, limited to long tweets (170 or more characters).

Classification problem	Precision[a]	Recall[a]	F1-score[a]
2	72.6	73.2	72.4

[a]These data are provided as percentages.

Figure 5. Confusion matrix for classification problem 2 limited to long tweets (n=168).

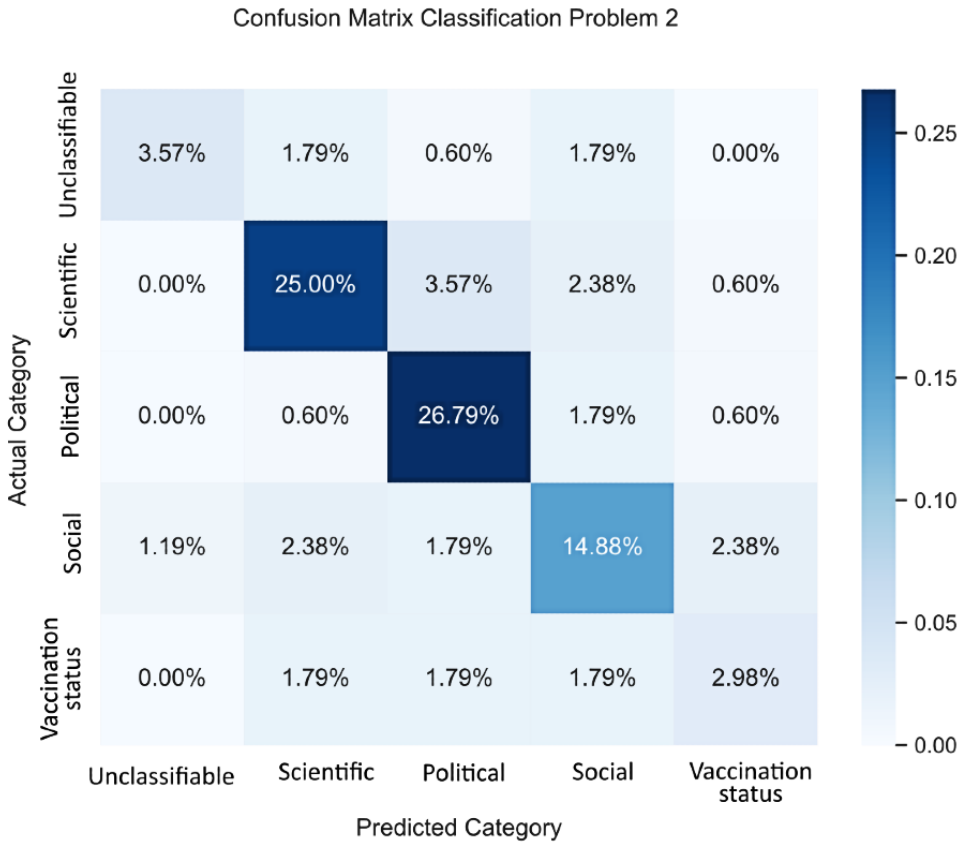


Table 7. The proportion of correct classifications for each label in classification problem 2, limited to long tweets (170 or more characters; n=168).

Type of problem	Number of tweets
Classification problem 2, n (%)	
Unclassifiable	6 (46.3)
Scientific	42 (79.2)
Political	45 (90)
Social	25 (65.8)
Vaccination status	5 (35.7)

Discussion

Principal Findings

A total of 2 types of classification were examined. The accuracy of the model was better with the second classification (67.6%; F1-score 62.9%) than the first classification (59%; F1-score 55.3%). This accuracy is slightly higher than that obtained by BERT for the same topic (vaccines) [17] and in the same range as previous findings [16,29]. However, CamemBERT obtained a better accuracy (78.7%-87.8%) in a study using dichotomous labels for tweets about eating disorders and using a preprocessing step, reducing the initial number of tweets by 2 [26]. However, by limiting the analysis to long tweets (170 or more characters, in accordance with the statistical analysis conducted on the performance of the model), the accuracy of classification model 2 improved significantly (from 62.9% to 72.4% for the F1-score).

Therefore, as shown by Kummervold et al [17], the classification choices have a significant influence on the accuracy of a model. As in other research areas, the vaccine hesitancy debate crystallizes the opposition. Individuals from the pro and con sides debated on Twitter after the announcement of the implementation of a health pass in July 2021 by the French president. The mobilized arguments were scientific or

pseudoscientific to justify or contest this political decision. Several Twitter users participated in the debate to convince anti-vaccine proponents to become vaccinated. Another group of users participated by joking about or ironizing the positions of each side.

Consequently, tweet content is so varied that it remains difficult to manually categorize, and this has been reflected in the model predictions. On the one hand, considering classification problem 1, tweets containing characteristic terms of the anti-vaccine position, such as "5G," "freedom," "phase of testing," "side effect," and "#passdelahonte" ("shameful pass"), were found to be easier to label and predict. However, because antivaccine proponents spread disinformation more widely on social media [30], the position of provaccine individuals is less polarized [7], which reduces the model's precision because the terms are less singular. On the other hand, considering classification problem 2, the classes were more distinctive since their lexical fields did not overlap. Indeed, when Twitter users commented on political decisions, the terminology used was different from that used to mobilize scientific or pseudoscientific arguments. Moreover, the scientific and political labels were best predicted by the model (67/84, 79.8% and 93/113, 82.3%, respectively).

Finally, relevant tweets for a topic may be rare in a data set. In some studies, the corpus is halved [13], while in others, only 0.5% (4000/810,600) of downloaded tweets were included in the analysis [16]. It would be interesting to find an objective method to improve model predictions without drastically reducing the data set. The approach of limiting tweet length can be an option, as we have demonstrated in this paper.

Limitations

Several limitations can be highlighted, including the following: (1) the data were only provided from a single social media platform (Twitter); (2) all tweets containing the term "vaccine" and its derivatives were included without preselection; (3) several categorization classes were unbalanced; (4) a larger training set could provide contrasting results; (5) the categorization choices could affect the performance of CamemBERT, as seen in the confusion matrix; and (6) the suggestions provided (limiting the number of tweet characters) may only apply to tweets on the topic of vaccination, so further studies are needed to confirm the relevance of our conclusions.

Conclusions

In this study, we tested the accuracy of a model (CamemBERT) without preselecting tweets, and we elaborated an epistemological reflection for future research. When the vaccine debate is jostled by contested political decisions, tweet content becomes so heterogeneous that the accuracy of the model decreases for the less differentiating classes. In summary, our analysis shows that epistemological choices (types of classes) can affect the accuracy of machine learning models. However, our tests also showed that it is possible to improve the model accuracy by using an objective method based on tweet length selection. Other possible avenues for improvement remain to be tested, such as the addition of features provided by Twitter (conservation ID, number of Twitter users following or followers, user public metrics listed count, user public metrics tweet count, or user ID).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Examples of conflicting labeling.
[PDF File (Adobe PDF File), 132 KB - medinform_v10i5e37831_app1.pdf]

References

1. Sauvayre R. Ethics of belief, trust and epistemic value: the case of the scientific controversy surrounding the measles vaccine. JLAE 2021 Sep 24;18(4):24-34. [doi: 10.33423/jlae.v18i4.4607]
2. Kata A. A postmodern Pandora's box: anti-vaccination misinformation on the Internet. Vaccine 2010 Feb 17;28(7):1709-1716. [doi: 10.1016/j.vaccine.2009.12.022] [Medline: 20045099]
3. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm--an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine 2012 May 28;30(25):3778-3789. [doi: 10.1016/j.vaccine.2011.11.112] [Medline: 22172504]
4. Aquino F, Donzelli G, De Franco E, Privitera G, Lopalco PL, Carducci A. The web and public confidence in MMR vaccination in Italy. Vaccine 2017 Aug 16;35(35 Pt B):4494-4498. [doi: 10.1016/j.vaccine.2017.07.029] [Medline: 28736200]
5. Deiner MS, Fathy C, Kim J, Niemeyer K, Ramirez D, Ackley SF, et al. Facebook and Twitter vaccine sentiment in response to measles outbreaks. Health Informatics J 2017 Nov 01:1460458217740723. [doi: 10.1177/1460458217740723] [Medline: 29148313]
6. Bavel JJV, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, et al. Using social and behavioural science to support COVID-19 pandemic response. Nat Hum Behav 2020 Apr 30:460-471. [doi: 10.1038/s41562-020-0884-z] [Medline: 32355299]
7. Sear RF, Velasquez N, Leahy R, Restrepo NJ, Oud SE, Gabriel N, et al. Quantifying COVID-19 content in the online health opinion war using machine learning. IEEE Access 2020;8:91886-91893 [FREE Full text] [doi: 10.1109/ACCESS.2020.2993967] [Medline: 34192099]

8. Kanozia R, Arya R. “Fake news”, religion, and COVID-19 vaccine hesitancy in India, Pakistan, and Bangladesh. Media Asia 2021 May 17;48(4):313-321. [doi: 10.1080/01296612.2021.1921963]
9. Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccin Immunother 2020 Jul 21:1-8. [doi: 10.1080/21645515.2020.1780846] [Medline: 32693678]
10. Edwards B, Biddle N, Gray M, Sollis K. COVID-19 vaccine hesitancy and resistance: correlates in a nationally representative longitudinal survey of the Australian population. PLoS One 2021;16(3):e0248892 [FREE Full text] [doi: 10.1371/journal.pone.0248892] [Medline: 33760836]
11. Antonakaki D, Fragopoulou P, Ioannidis S. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. Expert Syst Appl 2021 Feb;164:114006. [doi: 10.1016/j.eswa.2020.114006]
12. Hu T, Wang S, Luo W, Zhang M, Huang X, Yan Y, et al. Revealing public opinion towards COVID-19 vaccines with Twitter data in the United States: spatiotemporal perspective. J Med Internet Res 2021 Sep 10;23(9):e30854 [FREE Full text] [doi: 10.2196/30854] [Medline: 34346888]
13. Alshalan R, Al-Khalifa H, Alsaeed D, Al-Baity H, Alshalan S. Detection of hate speech in COVID-19-related Tweets in the Arab region: deep learning and topic modeling approach. J Med Internet Res 2020 Dec 08;22(12):e22609 [FREE Full text] [doi: 10.2196/22609] [Medline: 33207310]
14. D’Aniello G, Gaeta M, La Rocca I. KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. Artif Intell Rev 2022 Jan 09. [doi: 10.1007/s10462-021-10134-9]
15. Küçük D, Can F. Stance detection: a survey. ACM Comput Surv 2021 Jan 31;53(1):1-37. [doi: 10.1145/3369026]
16. Visweswaran S, Colditz JB, O'Halloran P, Han N, Taneja SB, Welling J, et al. Machine learning classifiers for Twitter surveillance of vaping: comparative machine learning study. J Med Internet Res 2020 Aug 12;22(8):e17478 [FREE Full text] [doi: 10.2196/17478] [Medline: 32784184]
17. 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]
18. Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie E, et al. CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 6-8, 2020; Online. [doi: 10.18653/v1/2020.acl-main.645]
19. Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, et al. FlauBERT: unsupervised language model pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference. FlauBERT: Unsupervised Language Model Pre-training for French. Proceedings of the 12th Language Resources and Evaluation Conference. 2020 May 13-15; Marseille, France. European Language Resources Association; 2020 Presented at: 12th Language Resources and Evaluation Conference; May 13-15, 2020; Marseille, France p. 2479-2490 URL: https://aclanthology.org/2020.lrec-1.302
20. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv 2019 (forthcoming) [FREE Full text] [doi: 10.48550/arXiv.1907.11692]
21. 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). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (LongShort Papers). 2019 June 2-7; Minneapolis, USA. Association for Computational Linguistics; 2019 Presented at: 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics; June 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/n18-2]
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: Thirty-first Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 4-9 URL: https://dl.acm.org/doi/pdf/10.5555/3295222.3295349
23. Twitter controller-to-controller data protection addendum. Twitter's General Data Protection Regulation Hub. URL: https://gdpr.twitter.com/en/controller-to-controller-transfers.html [accessed 2022-04-25]
24. Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of the 7th International Conference on Learning Representations. 2018 Presented at: Proceedings of the 7th International Conference on Learning Representations; May 6-9, 2019; New Orleans, LA URL: https://openreview.net/pdf?id=Bkg6RiCqY7
25. NLP-French-model-for-vaccine-tweets. GitHub. URL: https://github.com/cdchauvi/NLP-French-model-for-vaccine-tweets [accessed 2022-04-26]
26. Benítez-Andrades JA, Alija-Pérez J, Vidal M, Pastor-Vargas R, García-Ordás MT. Traditional machine learning models and bidirectional encoder representations from transformer (BERT)-based automatic classification of Tweets about eating disorders: algorithm development and validation study. JMIR Med Inform 2022 Feb 24;10(2):e34492 [FREE Full text] [doi: 10.2196/34492] [Medline: 35200156]

27. Shin D, Kam HJ, Jeon M, Kim HY. Automatic classification of thyroid findings using static and contextualized ensemble natural language processing systems: development study. JMIR Med Inform 2021 Sep 21;9(9):e30223 [FREE Full text] [doi: 10.2196/30223] [Medline: 34546183]
28. It's time to talk about ditching statistical significance. Nature 2019 Mar;567(7748):283. [doi: 10.1038/d41586-019-00874-8] [Medline: 30894740]
29. Kumar A, Singh JP, Dwivedi YK, Rana NP. A deep multi-modal neural network for informative Twitter content classification during emergencies. Ann Oper Res 2020 Jan 16. [doi: 10.1007/s10479-020-03514-x]
30. Germani F, Biller-Andorno N. The anti-vaccination infodemic on social media: a behavioral analysis. PLoS One 2021;16(3):e0247642 [FREE Full text] [doi: 10.1371/journal.pone.0247642] [Medline: 33657152]

Abbreviations

API: application programming interface
BERT: Bidirectional encoder representations from transformers
NLP: natural language processing
OR: odds ratio

Edited by T Hao; submitted 08.03.22; peer-reviewed by JA Benítez-Andrades, R Poluru, S Doan; comments to author 11.04.22; revised version received 01.05.22; accepted 04.05.22; published 17.05.22.

Please cite as:
Sauvayre R, Vernier J, Chauvière C
An Analysis of French-Language Tweets About COVID-19 Vaccines: Supervised Learning Approach
JMIR Med Inform 2022;10(5):e37831
URL: https://medinform.jmir.org/2022/5/e37831
doi:10.2196/37831
PMID:35512274

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org