JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 10 (2022), Issue 11 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Viewpoint

Original Papers

Corrigenda and Addenda

Review

Visit Types in Primary Care With Telehealth Use During the COVID-19 Pandemic: Systematic Review

Kanesha Ward[1], BClinSci, MRes; Sanjyot Vagholkar[2], MBBS (Hons), MPH, PhD; Fareeya Sakur[1], MPH; Neha Nafees Khatri[1], MBBS, MPH; Annie Y S Lau[1], BE, PhD

[1]Centre for Health Informatics, Australian Institute for Health Innovation, Macquarie University, North Ryde, Australia

[2]Primary Care, Faculty of Medicine, Health & Human Sciences, Macquarie University, North Ryde, Australia

Corresponding Author:
Kanesha Ward, BClinSci, MRes
Centre for Health Informatics
Australian Institute for Health Innovation
Macquarie University
75 Talavera Rd, North Ryde NSW
North Ryde, 2113
Australia
Phone: 61 48355552
Email: kanesha_ward@iinet.net.au

Abstract

Background: Telehealth was rapidly incorporated into primary care during the COVID-19 pandemic. However, there is limited evidence on which primary care visits used telehealth.

Objective: The objective of this study was to conduct a systematic review to assess what visit types in primary care with use of telehealth during the COVID-19 pandemic were reported; for each visit type identified in primary care, under what circumstances telehealth was suitable; and reported benefits and drawbacks of using telehealth in primary care during the COVID-19 pandemic.

Methods: This study was a systematic review using narrative synthesis. Studies were obtained from four databases (Ovid [MEDLINE], CINAHL Complete, PDQ-Evidence, and ProQuest) and gray literature (NSW Health, Royal Australian College of General Practitioners guidelines, and World Health Organization guidelines). In total, 3 independent reviewers screened studies featuring telehealth use during the COVID-19 pandemic in primary care. Levels of evidence were assessed according to the Grading of Recommendations Assessment, Development, and Evaluation. Critical appraisal was conducted using the Mixed Methods Appraisal Tool. Benefits and drawbacks of telehealth were assessed according to the National Quality Forum Telehealth Framework.

Results: A total of 19 studies, predominately cross-sectional surveys or interviews (13/19, 68%), were included. Seven primary care visit types were identified: *chronic condition management* (17/19, 89%), *existing patients* (17/19, 89%), *medication management* (17/19, 89%), *new patients* (16/19, 84%), *mental health/behavioral management* (15/19, 79%), *post–test result follow-up* (14/19, 74%), and *postdischarge follow-up* (7/19, 37%). Benefits and drawbacks of telehealth were reported across all visit types, with *chronic condition management* being one of the visits reporting the greatest *use* because of a pre-existing patient-provider relationship, established diagnosis, and lack of complex physical examinations. Both patients and clinicians reported benefits of telehealth, including improved convenience, focused discussions, and continuity of care despite social distancing. Reported drawbacks included technical barriers, impersonal interactions, and semi-established reimbursement models.

Conclusions: Telehealth was used for different visit types during the COVID-19 pandemic in primary care, with most visits for *chronic condition management*, *existing patients*, and *medication management*. Further research is required to validate our findings and explore the long-term impact of hybrid models of care for different visit types in primary care.

Trial Registration: PROSPERO CRD42022312202; https://tinyurl.com/5n82znf4

(JMIR Med Inform 2022;10(11):e40469) doi:10.2196/40469

KEYWORDS

primary care; general practitioners; telehealth; telemedicine; COVID-19; remote consultation; video consultation; eHealth

Introduction

Background

The COVID-19 pandemic has radically disrupted all aspects of health care, notably the rapid adaption of telehealth within routine care [1-3]. *Telehealth*, defined as telecommunications, videoconferencing, or other digital modes, is used to remotely deliver health-related services to patients [4,5]. Before the COVID-19 pandemic, telehealth provided convenience, specifically for patients living in rural or remote settings, but was not routinely used in health care settings [5]. Telehealth during the pandemic was used across many medical specialties such as internal medicine, psychiatry, preventative medicine, surgery, neurology, dermatology, pediatrics, and infectious diseases [6].

In particular, some general practitioners (GPs) and patients welcomed telehealth in primary care general practice settings during the pandemic. A survey conducted by the Royal Australian College of General Practitioners (RACGP) involving >420 Australian GPs saw 1 in 5 respondents report 61% to 80% of their patients requesting a telehealth consultation during the COVID-19 pandemic [7]. Some patients and GPs have advocated for the long-term use of telehealth beyond the COVID-19 pandemic, for example, in the form of hybrid models of care [1,7-9]. Several countries (eg, Australia, the United States, and the United Kingdom) have introduced long-term funding for telehealth in primary care because of the pandemic.

There is potential for telehealth in primary care in nonpandemic settings [1]. However, the current model of telehealth may not be fit to sustain the long-term delivery of primary care [2,10,11]. As the rapid adoption of telehealth and other forms of remote care is witnessed, its limitations need to be examined [10]. Most telehealth systems were rolled out rapidly without much research into the risks (eg, lack of patient choice, missed diagnoses, challenges to the patient-clinician relationship, and inequality experienced by those affected by the digital divide) [1,10,12]. Identifying which in-person encounters are *appropriate* to be supported by telehealth consultation is one of the critical questions facing today's health care delivery.

A cross-sectional study conducted by Donaghy et al [13] explored the acceptability and suitability of telehealth for specific encounters, where they reported telehealth as suitable for a range of patient visit types and concerns such as prescription refills, discussion-based activities, nonsensitive test results, and patients with chronic conditions with established diagnoses. A systematic review by Shah and Badawy [14] evaluated the feasibility, accessibility, satisfaction, and treatment outcomes related to telehealth services among pediatric populations, with findings suggesting telehealth to be a suitable alternative to in-person care. A previous systematic review by Snoswell et al [15] aimed to synthesize literature on the clinical effectiveness of telehealth for specific medical conditions from 2010 to 2019. However, to our knowledge, this is the first systematic review to focus on what visit types in primary care are suitable for telehealth based on studies where data were collected during the COVID-19 pandemic.

Objectives

The objective of this study was to conduct a systematic review to assess (1) what visit types in primary care with use of telehealth during the COVID-19 pandemic were reported; (2) for each visit type identified in primary care, under what circumstances telehealth was suitable; and (3) reported benefits and drawbacks of using telehealth in primary care during the COVID-19 pandemic.

Methods

Information Sources

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

Our search included the following electronic databases: Ovid (MEDLINE), CINAHL Complete, PDQ-Evidence, and ProQuest. Gray literature sources included NSW Health publications, RACGP guidelines, and World Health Organization guidelines.

Search Strategy

A modified population, exposure, and outcome [16] strategy was used, with *population* corresponding to primary care general practice clinicians and patients; *exposure* as the exposure to telehealth as a replacement of in-person consultation; and *outcomes* as benefits and drawbacks of telehealth, which are assessed according to the National Quality Forum (NQF) Telehealth Framework [17], namely, access to care, effectiveness, experience, and financial impact or cost. Clinical outcomes outside the scope outlined per the NQF telehealth measures were not analyzed in detail in this systematic review because of the lack of available data. However, clinical outcomes (eg, mental health status, shielding status, and number of examinations) were also extracted in Multimedia Appendix 2 [1,9,18-34] if they were available.

Individualized search strategies were formulated for each selected database with various Medical Subject Headings and searchable terms combined with Boolean operators. The complete search strategy is provided in Multimedia Appendix 3. An initial full search was conducted in March 2020. A final full search was conducted in August 2022. Conducting 2 searches ensured that the most recent and relevant literature was included in this systematic review analysis. Including both searches also reflects the rapid rate at which research is being conducted on telehealth services used in primary care settings following the COVID-19 pandemic.

Eligibility Criteria

Eligibility criteria were developed to include studies (1) published between December 2019 and August 2022 to encompass the COVID-19 era, (2) that discussed GP-patient consultations delivered within a telehealth format, (3) that provided insight into the visit types in primary care where telehealth was used, and (4) that included outcome measures on patients' or clinicians' perceived suitability of or satisfaction with the teleconsultation experience. Studies featuring multiple health care settings may also be included based on the fact that only data from primary care clinicians or patients were used for this systematic review.

The exclusion criteria were (1) telehealth services that did not reflect a consultation format (ie, did not involve bidirectional communication between clinician and patient) within the primary care general practice setting (specialist consultations excluded), (2) studies where it was not explicit in what visit type in primary care was telehealth being used, and (3) studies not written in English. The complete eligibility criteria are provided in Multimedia Appendix 4.

Article Selection Process

Initially, titles and abstracts of studies were retrieved using our search strategy and uploaded to an EndNote (Clarivate Analytics) library [35]. Duplicates were removed before uploading to the Rayyan software (Rayyan Systems Inc) [36] for titles and abstracts to be screened independently by three reviewers (KW, FS, and NNK). The full texts of the selected studies were assessed in greater detail by lead reviewer KW. Disagreements in article screening decisions were resolved through consensus.

Data Extraction and Management

Data from the included studies were extracted using an adapted version of the Joanna Briggs Institute data abstraction form (Multimedia Appendix 5) [37]. Publication details, study design, participant demographics, primary care visit type, telehealth intervention, and outcome measures were extracted from the included studies. Benefits and drawbacks of telehealth were extracted as outcome measures, presented according to the NQF Telehealth Framework. The NQF Telehealth Framework addresses the assessment of whether telehealth specifically can be used to deliver quality care and related outcomes in comparison with in-person consultations [16]. Definitions of each outcome measure used in this framework—namely, *access to care*, *effectiveness*, *experience*, and *financial impact or cost*—are reported in Textbox 1 [17]. Only relevant statistics or narrative excerpts were extracted. Effect measures were quoted from individual studies with no further statistical comparison.

Textbox 1. Outcome measures and their definitions according to the National Quality Forum Telehealth Framework.

Definitions of outcome measures

- Access to care: the ability to receive health services promptly and appropriately; consideration for accessibility to technology, living in rural and urban communities, living in medically underserved areas, access to appropriate health specialists, and provider capacity to provide care
- Effectiveness: the systematic, clinical, operational, and technical success or barriers of telehealth; considerations of the overall system and care coordination established, impact on health outcomes or quality, how clinically integrated telehealth is within the health center, and ability to record and transmit necessary data
- Experience: the usability and effect of telehealth on patients and providers with consideration of the appropriateness of services, increase in patients' knowledge of care, patient compliance with care regimens, the difference in morbidity and mortality rates, patient safety, patient-centeredness, efficiency, diagnostic accuracy, ability to obtain actionable information, comfort, and satisfaction
- Financial impact or cost: potential cost savings or losses to patients, families, or providers regarding costs to access care, travel expenses, added value, and feasibility surrounding the technology involved

Critical Appraisal of the Included Studies

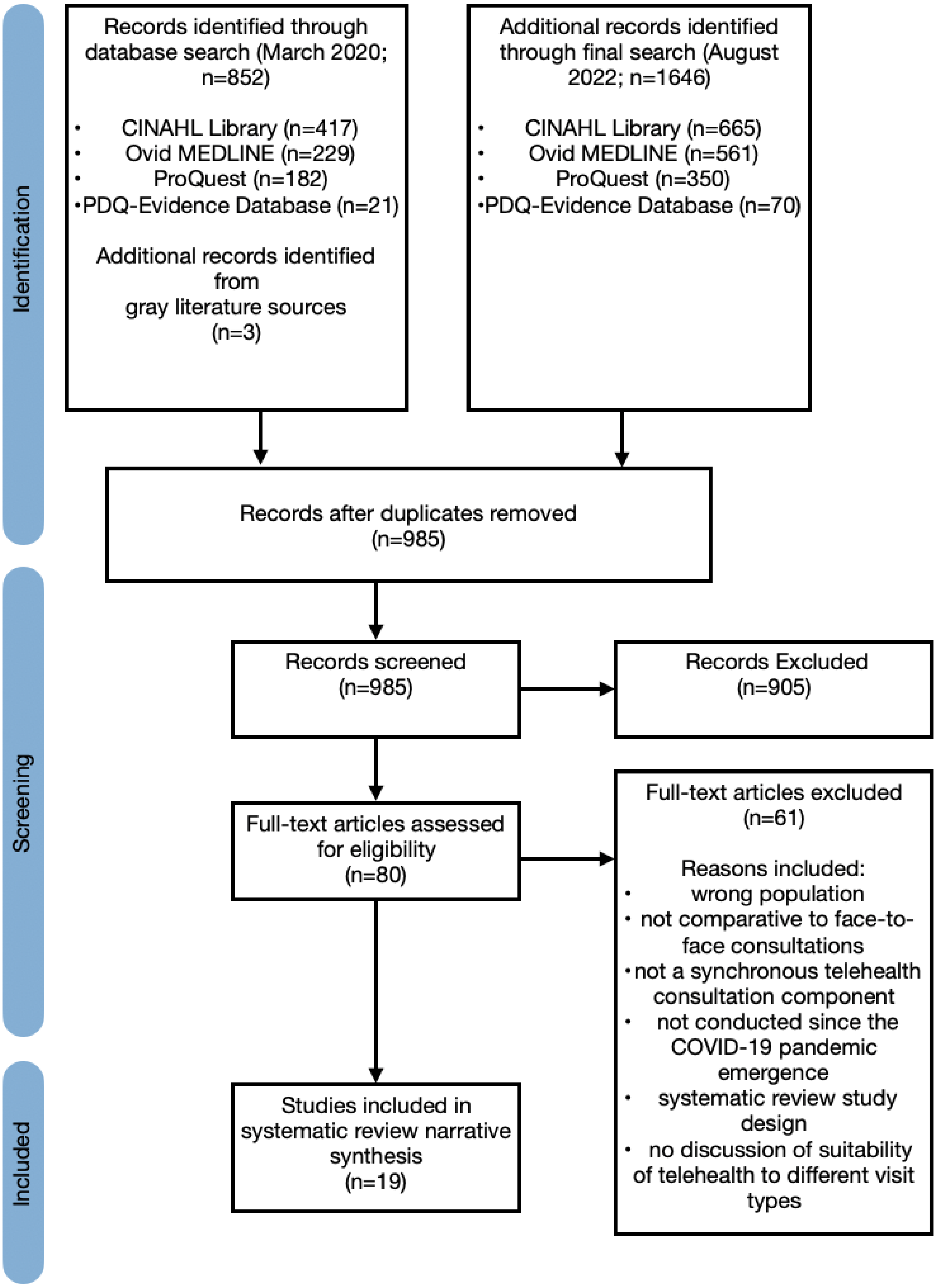
One reviewer (KW) led the critical appraisal. The Mixed Methods Appraisal Tool was used to appraise study designs of qualitative, quantitative, and mixed methods studies [38]. The level of evidence was assessed according to the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) [39]. Studies were not excluded based on outcomes of the critical appraisal; however, it was used to interpret findings. More details of the critical appraisal are provided in Multimedia Appendix 6 [39,40].

Results

Screening Process

Figure 1 outlines the article screening process, where 19 studies met the eligibility criteria and were included in a narrative synthesis.

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


Identification
Screening
Included
Records identified through database search (March 2020; n=852)
• CINAHL Library (n=417)
• Ovid MEDLINE (n=229)
• ProQuest (n=182)
•PDQ-Evidence Database (n=21)
Additional records identified from gray literature sources (n=3)
Additional records identified through final search (August 2022; n=1646)
• CINAHL Library (n=665)
• Ovid MEDLINE (n=561)
• ProQuest (n=350)
•PDQ-Evidence Database (n=70)
Records after duplicates removed (n=985)
Records screened (n=985)
Records Excluded (n=905)
Full-text articles assessed for eligibility (n=80)
Full-text articles excluded (n=61)
Reasons included:
• wrong population
• not comparative to face-to-face consultations
• not a synchronous telehealth consultation component
• not conducted since the COVID-19 pandemic emergence
• systematic review study design
• no discussion of suitability of telehealth to different visit types
Studies included in systematic review narrative synthesis (n=19)

Study Characteristics

Of the 19 included studies, 6 (32%) were conducted in the United States; 4 (21%) were conducted in the United Kingdom; 7 (37%) were conducted in Europe (Norway, Germany, Sweden, Netherlands, and Denmark); 3 (16%) were conducted in the Middle East (Israel and Oman); and the remaining 8 (42%) were conducted in Pakistan, Australia, and New Zealand (Multimedia Appendix 2).

Telephone communication (17/19, 89%) was the most frequent telehealth intervention in our included studies, followed by video communication (15/19, 79%), SMS text messaging (6/19, 32%), and email messaging (6/19, 32%). Table 1 provides a statistical breakdown of the types of telehealth interventions in the included studies.

Table 1. A statistical breakdown of the types of telehealth modes in the included studies (n=19).

Type of telehealth mode[a]	Studies, n (%)
Telephone communication	17 (89)
Video communication	15 (79)
SMS text messaging	6 (32)
Email messaging	6 (32)

[a]The included studies can discuss more than one telehealth mode.

Visit Types in Primary Care With Telehealth Support During the COVID-19 Pandemic That Were Reported

Visit types in primary care with telehealth support during the COVID-19 pandemic that were reported are outlined in Textbox 2. Definitions of each visit type were informed by Medicare item descriptions (Multimedia Appendix 7 [29,41-43]) after extraction from the included studies.

Table 2 and Table 3 outline the reported benefits and drawbacks of using telehealth during the COVID-19 pandemic for each visit type in primary care. Seven visit types in primary care with telehealth use during the COVID-19 pandemic were reported, namely, *chronic condition management* (17/19, 89%), *existing patients* (17/19, 89%), *medication management* (17/19, 89%), *new patients* (16/19, 84%), *mental health/behavioral management* (15/19, 79%), *post–test result follow-up* (14/19, 74%), and *postdischarge follow-up* (7/19, 37%).

Textbox 2. Visit types in primary care with telehealth support during the COVID-19 pandemic that were reported [41]. Visit types do not categorize within age groups. Patient age is considered as a benefit or drawback finding for this review.

Visit types and description

- Chronic condition management: 6-month or other routine chronic condition reviews, diabetes checkups, asthma or chronic obstructive pulmonary disease medication or management reviews, or chronic pain (ie, arthritis or musculoskeletal pain) discussions
- Mental health and behavioral management: anxiety, depression, behavioral treatment reviews, talking therapy, or mental health medication reviews; specialist visits excluded from this review
- Medication management: acute concerns (ie, antibiotics), medication reviews, oral contraceptive prescriptions, or dermatology prescriptions
- Post–test result follow-up: follow-up after magnetic resonance imaging examinations, x-rays, blood tests, or laboratory testing with their general practitioner (GP) to discuss given results
- Postdischarge follow-up: follow-up after a procedure or discharge from the hospital for patients with cancer after tumor removals, hospital admission following acute severe adverse reaction, or after pregnancy delivery
- Existing patients (acute or existing concerns): standard consultations with an annual checkup session or acute concerns (ie, cold or flu symptoms or dermatology concerns) with a patient the GP has a pre-existing patient-provider relationship; inclusive of patients with COVID-19 or shielding patients
- New patients (acute or existing concerns): standard consultations such as one-off sessions (eg, vaccination) or acute concerns (ie, cold or flu symptoms or dermatology concerns) with a patient with whom the GP has no pre-existing patient-provider relationship; inclusive of patients with COVID-19 or shielding patients

Table 2. Reported general practitioner-patient visit types with telehealth support during the COVID-19 pandemic (N=19).[a]

Visit type	Studies that reported the use of telehealth	Benefit findings of telehealth					
		Studies, n (%)	Level 1	Level 2	Level 3	Level 4	Level 5
Chronic condition management (n=17)	Johnsen et al [1] De Guzman et al [9] Jetty et al [18] Jabbarpour et al [20] Van de Poll-Franse et al [21] Gomez et al [22] Hasani et al [23] Imlach et al [24] Gabrielsson-Jarhult et al [25] Murphy et al [26] Schweiberger et al [27] RACGP[b] [28] MBS[c] [29] Mozes et al [30] Javanparast et al [31] Assing Hvidt et al [32] Due et al [33]	13 (76)	N/A[d]	N/A	Murphy et al [26]	Johnsen et al [1] De Guzman et al [9] Jetty et al [18] Jabbarpour et al [20] Gomez et al [22] Hasani et al [23] Schweiberger et al [27] Javanparast et al [31] Assing Hvidt et al [32] Due et al [33]	RACGP [28] MBS [29]
Medication management (nonchronic condition; n=17)	Johnsen et al [1] De Guzman et al [9] Jetty et al [18] Jabbarpour et al [20] Van de Poll-Franse et al [21] Gomez et al [22] Hasani et al [23] Imlach et al [24] Gabrielsson-Jarhult et al [25] Murphy et al [26] Schweiberger et al [27] RACGP [28] MBS [29] Mozes et al [30] Javanparast et al [31] Assing Hvidt et al [32] Due et al [33]	11 (65)	N/A	N/A	Gabrielsson-Jarhult et al [25]	Johnsen et al [1] De Guzman et al [9] Jetty et al [18] Gomez et al [22] Hasani et al [23] MBS [29] Mozes et al [30] Due et al [33]	RACGP [28] MBS [29]
Existing patients (acute or existing concern; n=17)	Johnsen et al [1] De Guzman et al [9] Jetty et al [18] Grossman et al [19] Jabbarpour et al [20] Van de Poll-Franse et al [21] Gomez et al [22] Hasani et al [23] Imlach et al [24] Murphy et al [26] Schweiberger et al [27] RACGP [28] MBS [29] Mozes et al [30] Javanparast et al [31] Assing Hvidt et al [32] Manski-Nankervis et al [34]	11 (65)	N/A	N/A	Imlach et al [24] Murphy et al [26] Mozes et al [30]	Johnsen et al [1] De Guzman et al [9] Grossman et al [19] Hasani et al [23] Schweiberger et al [27] RACGP [28] MBS [29] Assing Hvidt et al [32]	—[e]

Visit type	Studies that reported the use of telehealth	Benefit findings of telehealth					
		Studies, n (%)	Level 1	Level 2	Level 3	Level 4	Level 5
New patients (acute or existing concern; n=16)	• Johnsen et al [1] • De Guzman et al [9] • Jetty et al [18] • Grossman et al [19] • Jabbarpour et al [20] • Van de Poll-Franse et al [21] • Gomez et al [22] • Hasani et al [23] • Imlach et al [24] • Gabrielsson-Jarhult et al [25] • Schweiberger et al [27] • RACGP [28] • MBS [29] • Assing Hvidt et al [32] • Due et al [33] • Manski-Nankervis et al [34]	7 (44)	N/A	N/A	• Gabrielsson-Jarhult et al [25]	• Johnsen et al [1] • Hasani et al [23] • Schweiberger et al [27] • Assing Hvidt et al [32] • Due et al [33]	• MBS [29]
Mental health and behavioral management (n=15)	• Johnsen et al [1] • De Guzman et al [9] • Jetty et al [18] • Jabbarpour et al [20] • Gomez et al [22] • Hasani et al [23] • Imlach et al [24] • Murphy et al [26] • Schweiberger et al [27] • RACGP [28] • MBS [29] • Javanparast et al [31] • Assing Hvidt et al [32] • Due et al [33] • Manski-Nankervis et al [34]	11 (73)	N/A	N/A	• Imlach et al [24] • Murphy et al [26]	• Johnsen et al [1] • De Guzman et al [9] • Jabbarpour et al [20] • Hasani et al [23] • Schweiberger et al [27] • Assing Hvidt et al [32] • Due et al [33]	• RACGP [28] • MBS [29]
Post–test result follow-up (n=14)	• Johnsen et al [1] • De Guzman et al [9] • Jetty et al [18] • Jabbarpour et al [20] • Van de Poll-Franse et al [21] • Gomez et al [22] • Hasani et al [23] • Imlach et al [24] • Murphy et al [26] • Schweiberger et al [27] • RACGP [28] • MBS [29] • Assing Hvidt et al [32] • Due et al [33]	5 (36)	N/A	N/A	—	• Johnsen et al [9] • Gomez et al [22] • Hasani et al [23] • Assing Hvidt et al [32] • Due et al [33]	—
Postdischarge follow-up (n=7)	• Johnsen et al [1] • Jetty et al [18] • Hasani et al [23] • Imlach et al [24] • Murphy et al [26] • Mozes et al [30] • Javanparast et al [31]	5 (71)	N/A	N/A	• Murphy et al [26]	• Hasani et al [23]	• RACGP [28] • MBS [29]

[a]Definitions of the different visit types are informed by the Department of Health Medicare Benefits Scheme item definitions [41] (Multimedia Appendix 7). Levels of evidence were derived from the Grading of Recommendations Assessment, Development, and Evaluation scoring [39]. Level 1 is systematic reviews, level 2 is randomized controlled trials, level 3 is nonrandomized experimental studies or comparative (observational) studies, level 4 is case series (cohort studies), and level 5 is opinion pieces or clinical guidelines. Each article can report more than one visit type supported with telehealth during the COVID-19 pandemic.

[b]RACGP: Royal Australian College of General Practitioners.

[c] MBS: Medicare Benefits Schedule.

[d] N/A: not applicable.

[e] No data available for the category specified.

Table 3. Reported general practitioner-patient visit types with drawback findings of telehealth during the COVID-19 pandemic (N=19).[a]

Visit type	Studies that reported the use of telehealth	Drawback findings of telehealth					
		Studies, n (%)	Level 1	Level 2	Level 3	Level 4	Level 5
Chronic condition management (n=17)	• Johnsen et al [1] • De Guzman et al [9] • Jetty et al [18] • Jabbarpour et al [20] • Van de Poll-Franse et al [21] • Gomez et al [22] • Hasani et al [23] • Imlach et al [24] • Gabrielsson-Jarhult et al [25] • Murphy et al [26] • Schweiberger et al [27] • RACGP[b] [28] • MBS[c] [29] • Mozes et al [30] • Javanparast et al [31] • Assing Hvidt et al [32] • Due et al [33]	6 (35)	N/A[d]	N/A	• Gabrielsson-Jarhult et al [25] • Mozes et al [30]	• Johnsen et al [1] • De Guzman et al [9] • Van de Poll-Franse et al [21] • Due et al [33]	—[e]
Medication management (nonchronic condition; n=17)	• Johnsen et al [1] • De Guzman et al [9] • Jetty et al [18] • Jabbarpour et al [20] • Van de Poll-Franse et al [21] • Gomez et al [22] • Hasani et al [23] • Imlach et al [24] • Gabrielsson-Jarhult et al [25] • Murphy et al [26] • Schweiberger et al [27] • RACGP [28] • MBS [29] • Mozes et al [30] • Javanparast et al [31] • Assing Hvidt et al [32] • Due et al [33]	3 (18)	N/A	N/A	• Imlach et al [24] • Mozes et al [30]	• Johnsen et al [1]	—
Existing patients (acute or existing concern; n=17)	• Johnsen et al [1] • De Guzman et al [9] • Jetty et al [18] • Grossman et al [19] • Jabbarpour et al [20] • Van de Poll-Franse et al [21] • Gomez et al [22] • Hasani et al [23] • Imlach et al [24] • Murphy et al [26] • Schweiberger et al [27] • RACGP [28] • MBS [29] • Mozes et al [30] • Javanparast et al [31] • Assing Hvidt et al [32] • Manski-Nankervis et al [34]	1 (6)	N/A	N/A	—	• De Guzman et al [9]	—

Visit type	Studies that reported the use of telehealth	Drawback findings of telehealth					
		Studies, n (%)	Level 1	Level 2	Level 3	Level 4	Level 5
New patients (acute or existing concern; n=16)	- Johnsen et al [1] - De Guzman et al [9] - Jetty et al [18] - Grossman et al [19] - Jabbarpour et al [20] - Van de Poll-Franse et al [21] - Gomez et al [22] - Hasani et al [23] - Imlach et al [24] - Gabrielsson-Jarhult et al [25] - Schweiberger et al [27] - RACGP [28] - MBS [29] - Assing Hvidt et al [32] - Due et al [33] - Manski-Nankervis et al [34]	9 (56)	N/A	N/A	—	- Johnsen et al [1] - De Guzman et al [9] - Gomez et al [22] - Hasani et al [23] - Gabrielsson-Jarhult et al [25] - Assing Hvidt et al [32] - Due et al [33]	- RACGP [28] - MBS [29]
Mental health and behavioral management (n=15)	- Johnsen et al [1] - De Guzman et al [9] - Jetty et al [18] - Jabbarpour et al [20] - Gomez et al [22] - Hasani et al [23] - Imlach et al [24] - Murphy et al [26] - Schweiberger et al [27] - RACGP [28] - MBS [29] - Javanparast et al [31] - Assing Hvidt et al [32] - Due et al [33] - Manski-Nankervis et al [34]	3 (20)	N/A	N/A	—	- De Guzman et al [9] - Due et al [33] - Manski-Nankervis et al [34]	—
Post–test result follow-up (n=14)	- Johnsen et al [1] - De Guzman et al [9] - Jetty et al [18] - Jabbarpour et al [20] - Van de Poll-Franse et al [21] - Gomez et al [22] - Hasani et al [23] - Imlach et al [24] - Murphy et al [26] - Schweiberger et al [27] - RACGP [28] - MBS [29] - Assing Hvidt et al [32] - Due et al [33]	3 (21)	N/A	N/A	—	- Jetty et al [18] - Hasani et al [23] - Due et al [33]	—
Postdischarge follow-up (n=7)	- Johnsen et al [1] - Jetty et al [18] - Hasani et al [23] - Imlach et al [24] - Murphy et al [26] - Mozes et al [30] - Javanparast et al [31]	3 (43)	N/A	N/A	—	- Johnsen et al [1] - Jetty et al [18] - Hasani et al [23]	—

[a] Definitions of the different visit types are informed by the Department of Health Medicare Benefits Scheme item definitions [41] (Multimedia Appendix 7). Levels of evidence were derived from the Grading of Recommendations Assessment, Development, and Evaluation scoring [39]. Level 1 is systematic reviews, level 2 is randomized controlled trials, level 3 is nonrandomized experimental studies or comparative (observational) studies, level 4 is case series (cohort studies), and level 5 is opinion pieces or clinical guidelines. Each article can report more than one visit type supported with telehealth during the COVID-19 pandemic.

[b] RACGP: Royal Australian College of General Practitioners.

[c] MBS: Medicare Benefits Schedule.

[d] N/A: not applicable.

[e] No data available for the category specified.

The benefits and drawbacks of using telehealth during the COVID-19 pandemic in primary care were reported across all visit types. Visit types with >60% of studies reporting benefits included *chronic condition management*, *mental health/behavioral management*, *medication management*, and *existing patients*, whereas the visit types with 40% of studies reporting drawbacks of telehealth included *new patients* and *postdischarge follow-up*.

Diverse study designs according to GRADE were reported in the included studies, with most (13/19, 68%) corresponding to level-4 evidence (cohort studies, interviews, and surveys), followed by level 3 (nonrandomized experimental studies or comparative or observational studies; 4/19, 21%) and level 5 (opinion pieces or clinical guidelines; 2/19, 11%). No randomized controlled trials (level-2 evidence) or systematic reviews (level-1 evidence) were found to have met the eligibility criteria to be included in this systematic review.

Suitability of Using Telehealth Support for Each Visit Type During the COVID-19 Pandemic

Overview

For each visit type in primary care during the COVID-19 pandemic where telehealth support was reported, benefits and drawbacks are outlined in this section. Table 4 provides a summary of the circumstances when telehealth was reported as suitable and not suitable per patient visit types during the COVID-19 pandemic. For more details on supporting evidence, please refer to Multimedia Appendix 8 [1,9,20-26,28-34, 39,40,44-48].

Table 4. Circumstances when telehealth was reported as suitable and not suitable per patient visit types during the COVID-19 pandemic.

Visit type and subcategory	Circumstances when telehealth was suitable	Circumstances when telehealth was NOT suitable
Condition- or concern-based		
Chronic condition management	• Pre-existing patient-provider relationship [1,23,27] • Established diagnosis [18] • Lack of complex physical examinations [20]	• Chronic conditions when there were complex issues requiring close monitoring or longer consultations (eg, complex comorbidities, cancer, complex social issues, low hearing and vision, and cognitive impairment) [1,9,25,30]
Medication (nonchronic condition) management	• Prescription refills of existing medications [1,9,22,32] • Simple, straightforward health concerns (eg, oral contraceptives) [1,22] • Predominately discussion-based activities [1,22]	• When physical examinations were necessary (eg, prescribing antibiotics) [1,24,31] • Prescription of new medications [1,24]
Mental health and behavioral management	• Patients with mild mental health issues (ie, not at risk to themselves or others or without high cognitive impairments) [20] • Patients who did not prefer a physical presence [9,20] • Predominately discussion-based and counseling activities [1,9,20,23,33]	• When cultural, language, or confidentiality concerns affected patients' ability to communicate or disclose [20,26] • Patients with unstable mental health concerns (eg, suicidal ideation) [1] • When physical examinations were necessary for screening tests or psychotherapy delivery [1]
Post–test result follow-up	• Predominately discussion tasks rather than physical examinations [22,23,26] • When patients preferred to view test results via video compared with in person [26] • Nonsensitive test results [9]	• When discussing sensitive test results (eg, positive cancer diagnosis) [33] • When explaining complex medical jargon used in test results [33]
Postdischarge follow-up	• When patients lived far away or had difficulty arranging a same-day visit or frequent follow-ups [23,26] • Patients with pre-existing patient-provider relationships at the postoperative clinic [1]	• When complex physical examinations were needed [18,23] • When multiple care team members (eg, nurses) were needed to address physical aspects of care (eg, wound care) [23]
Patient characteristics–based		
Existing patients (acute or existing concern)	• Pre-existing patient-provider relationship [1,23,24,27] • Established understanding of patients' history [1,23,24,27] • Pre-established rapport [1,23,24,27] • Issues primarily reliant on assessing visual symptoms (eg, dermatological concerns) [32]	• New diagnoses even with pre-existing patient-provider relationships [9] • Severe concern that required more physical examinations (eg, chest pain or stomach pain) [30]
New patients (acute or existing concern)	• New patients when the consultation focused on pre-existing diagnosed concerns [1,23,25,27] • Simple acute concerns (eg, dermatological concerns) that could be assessed using photos or video without complex physical examinations [1,23,25,27]	• New diagnoses with no pre-existing patient-provider relationship or lack of knowledge of patient history [1,22] • New patients with difficult or complex symptoms that relied on self-reported information or self-examinations [1,22] • When patients were not forthcoming (eg, shyness or language or cultural barriers) [1,22] • Technical issues affecting building rapport [33]

Chronic Condition Management

Chronic condition management visits in primary care were reported as being one of the visit types with the greatest number of studies reporting the use of telehealth during the COVID-19 pandemic (17/19, 89%). Of these 17 studies, 13 (76%) reported benefits and 6 (35%) reported drawbacks.

In the included studies, chronic condition management visits were often reported as suitable for telehealth because of *a pre-existing patient-provider relationship*, *established diagnosis*, *and lack of complex physical examinations*. Routine visits for chronic conditions (eg, diabetes checkups) could be facilitated using telehealth, as these tasks often relied on discussions (eg, diet and medication) [9,23,33]. Some of the examinations could be completed by patients at home under clinician guidance, such as foot examinations or weight measurements [1,18,20,22,27]. Patients could show their list of medications at home by reading the labels [22], and they could be educated on ways to use and administer medications at home (eg, asthma inhalers) and assisting with potential safety hazards or home support systems (eg, pets) [23]. Self-management education could also be enhanced if patients could share with their GPs during telehealth their home setting and at-home tools (eg, at-home blood pressure cuffs, glucose monitors, and heart rate monitors) [22].

A drawback of telehealth for chronic condition management in the included studies was when close monitoring (eg, complex comorbidities or cancer diagnoses) [30] or complex physical examinations (eg, pediatric examinations or smear examinations) were required [33]. Some patients reported being hesitant to use telehealth because of their unfamiliarity with the technology [21]. However, most patients with chronic conditions expressed high satisfaction and willingness to engage with telehealth again [31]. Patients with chronic conditions particularly favored the remote nature of telehealth as they were often at a higher risk of adverse symptoms if infected with COVID-19 when attending in-person clinics [21,23,28,29].

Existing Patients

Existing patient consultations were reported as being one of the visit types with the greatest number of studies reporting the use of telehealth during the COVID-19 pandemic (17/19, 89% of the included studies). Of these 17 studies, 11 (65%) reported benefits and 1 (6%) reported drawbacks.

In the included studies, telehealth was reported as suitable for visits with a pre-existing patient-provider relationship as clinicians understood the patients' history and had a pre-established rapport [1,23,24,27,32]. Issues primarily reliant on assessing visual symptoms, such as dermatological concerns, could be shared with clinicians via photos or video [32]. In some cases, clinicians reported higher efficiency using telehealth. They could reduce the downtime involved in transiting between different patients during in-person encounters and see more patients via telehealth [9].

In the included studies, existing patients reported satisfaction with telehealth, especially for straightforward matters (eg, medication refill) and patients at high risk of COVID-19 [24,26,30]. However, a drawback of telehealth was when new diagnoses were involved, even among people with pre-existing patient-provider relationships, because of the poor ability to conduct physical examinations [9].

Medication Management

Medication management consultations were reported as being one of the visit types with the greatest number of studies reporting the use of telehealth during the COVID-19 pandemic (17/19, 89% of eligible studies). Of these 17 studies, 11 (65%) reported benefits and 3 (18%) reported drawbacks.

In the included studies, telehealth was reported as making medication reconciliations easier, improving patients' adherence to their medications [1,9,22,32]. Telehealth was reported as supporting prescription refills for patients familiar with the medication's side effects and risks and for straightforward health concerns such as oral contraceptives [9,31,32]. Patients reported being satisfied with their telehealth experience related to medications [1,18,25]. For example, patients could share their medications at home via video and image sharing with their clinicians. Furthermore, clinicians expressed greater relief when not being pressured to prescribe addictive drugs to at-risk patients during telehealth [22].

Drawbacks reported in the studies included concerns when physical examinations were necessary (eg, checking for infections when prescribing antibiotics) and prescription of new medications [1,24,30]. Poorer communication in patient education of medications was also observed in some telehealth consultations, potentially affecting patients' understanding of their medications [1,24].

New Patients

New patient consultations with the use of telehealth during the COVID-19 pandemic were reported in 84% (16/19) of eligible studies. Of these 16 studies, 7 (44%) reported benefits and 9 (56%) reported drawbacks.

In the included studies, telehealth was reported as only suitable for new patients when the consultation focused on pre-existing diagnosed concerns, acute concerns (eg, dermatological concerns) that could be assessed via visual cues (such as via photo or video sharing), or when there was no need for physical examinations [1,23,25,27]. It is important to note that new patients are not always supported by health care reimbursement (eg, Medicare for Australian patients) outside certain criteria (ie, positive COVID-19 status, close contact, hot spot area, and emergency consultation), which affects the number of studies included for this visit type.

However, telehealth was reported as not suitable for new diagnoses when there was no pre-existing patient-provider relationship, lack of knowledge of patient history, or no pre-established patient rapport [1,22]. Telehealth for new patients would be particularly difficult when complex symptoms are involved or when patients are not forthcoming with their concerns (eg, feeling shy or experiencing language or cultural barriers). Managing new patients over telehealth would rely on trusting patients' self-reported symptoms and patient-directed examination, which can become complicated when there is an absence of pre-existing knowledge of the patient [32]. In addition, technical problems within telehealth consultations can make building rapport with new patients even harder [33].

Mental Health and Behavioral Management

Mental health and behavioral management consultations with the use of telehealth during the COVID-19 pandemic were reported in 79% (15/19) of the included studies. Of these 15 studies, 11 (73%) reported benefits and 3 (20%) reported drawbacks.

In the included studies, telehealth was reported as only suitable for mental health and behavioral management when the consultation predominately focused on discussion and counseling activities [1,20,23,33]. Telehealth was suitable for patients with mild mental health issues (ie, patients not at risk to themselves or others), those without high cognitive impairments, and those who did not prefer a physical presence in consultations [20]. Studies involving patients with more complex mental health concerns referred to specialists (ie, psychiatrists) and participants in specialist mental health telehealth programs were excluded from this review. Patients with mental health concerns reported the benefits of reduced wait times when using telehealth during the COVID-19 pandemic, resulting in fewer barriers to accessing mental health care support [26,27]. Patients also reported being satisfied with their telehealth experience for mental health issues during the COVID-19 pandemic, particularly because of consultations being completed in a discreet manner (ie, the privacy of their own home) [1,9,24].

Drawbacks reported in the studies included patients' hesitancy to disclose over telehealth because of the stigma around mental health concerns, cultural or language barriers, and confidentiality around disclosing sensitive matters where there was a lack of privacy at home [20,26]. It was challenging to conduct telehealth consultations with patients with unstable mental health concerns (eg, suicidal ideation) [1] or concerns requiring lengthier consultations [9]. There were mixed views about the need for physical examinations for screening tests [31].

Post–Test Result Follow-up

Post–test result follow-up consultations with the use of telehealth during the COVID-19 pandemic were reported in 74% (14/19) of the included studies. Of these 14 studies, 5 (36%) reported benefits and 3 (21%) reported drawbacks.

In the included studies, post–test result follow-up was often suitable for telehealth as the primary activity involved discussions of test results rather than conducting physical examinations [22,23,26,32,33]. For patients and clinicians, practices used procedures to ensure confidentiality via telehealth when receiving (and discussing) test results [23]. For example, some practices used confirmation ID numbers or asked patients to confirm their date of birth before revealing sensitive medical information because of the absence of in-person confirmation [23]. In some cases, telehealth also improved the ability to share test results with patients compared with in-person consultations (eg, screen sharing of test results with patients over video consultation) rather than the patient attempting to reach over to read the test result on the GP's computer screen during in-person encounters [26].

A drawback of telehealth reported for this visit type was the poorer communication patterns observed when explaining to patients complex medical jargon used in test results. This is possibly related to the impersonal nature of telehealth, the inability to use visual aids, the lack of a physical presence, or other elements required to explain test results remotely [33]. Other clinic staff may communicate test results if results are satisfactory or do not require additional follow-up, resulting in minimal benefit and drawback findings reported for this visit type. Unsatisfactory results may lead to GP-patient consultations, possibly resulting in more drawbacks reported for this visit type.

Postdischarge Follow-up

Postdischarge follow-up consultations with the use of telehealth during the COVID-19 pandemic were reported in 37% (7/19) of the included studies. Of these 7 studies, 5 (71%) reported benefits and 3 (43%) reported drawbacks.

In the included studies, telehealth was reported as suitable for postdischarge follow-up visits when patients lived far away or had difficulty arranging same-day or frequent follow-up visits (eg, antenatal visits) [23,26]. Patients with pre-existing patient-provider relationships linked to the same postoperative clinic also reported satisfaction [1].

However, there was the drawback of it being harder to coordinate care [1]. This visit type often involved multiple care team members and complex physical examinations by various clinic members (eg, nurses and practitioners to address wound care) [18,23]. It was also challenging to share documentation from multiple team members [23].

Benefits and Drawbacks of Using Telehealth in Primary Care During the COVID-19 Pandemic

This section outlines the benefits and drawbacks of using telehealth in primary care during the COVID-19 pandemic from patient and clinician perspectives reported according to the NQF Telehealth Framework. For more details on supporting evidence, please refer to Multimedia Appendix 9 [1,9,19,21-34].

Access to Care

The NQF outcome measure "Access to care" (ie, the ability to receive health services promptly and appropriately) was reported in 84% (16/19) of the included studies. A summary of benefits and drawbacks of telehealth per this outcome factor is provided in Table 5.

Both patients and clinicians reported the benefits of using telehealth to maintain timely and frequent contact, shortening wait times in between visits and having a satisfactory experience [1,22,25,26,30,32]. Patients particularly enjoyed the additional benefits of reduced travel time [32,33], the convenience of being at home [23,31], having quicker access to care for simple concerns [1,26,29], and being able to access care that was only available for a teleconsultation but not available for in-person consultations (eg, outside clinic opening hours) [19,30], whereas clinicians reported the benefits of seeing more patients using telehealth [28] and connecting with patients who preferred technology over in-person encounters [9,25].

Table 5. Benefits and drawbacks of telehealth according to the “Access to Care” outcome factor per perspective.

Perspective and benefits of telehealth	Drawbacks of telehealth
Primary care clinician perspective	
• Greater number of patients that can be seen using telehealth compared with in person (ie, teleconsultations tend to be shorter and more convenient, reducing cancelation rates) [25,26] • Enables clinicians to connect with patients who may prefer technology over in-person encounters [25]	• Harder to address language or cognition barriers [32] • Need to address risks associated with digital platforms (eg, cyberattacks, security, and confidentiality in web-based communication) [25]
Patient perspective	
• Reduced travel time [31,34] • Improved convenience [1,22,25,26,30,31] • Ability to book consultations outside clinic hours [25,30] • Ability to access care quicker owing to not requiring the same clinician for simple concerns [25,31,34]	• Excludes and deters potentially at-risk patients who are not familiar with the technology [21,22]
Both primary care clinician and patient perspective	
• Satisfied with access and technical quality in most telehealth consultations [1,18] • Timely and more frequent access to care for at-risk patients because of convenience and shortened wait times [1,26,27,30]	• Insufficient technical support, infrastructure, or equipment to access telehealth [33] • Varying complexity of telehealth systems needed because of different complexities in patients’ health conditions (eg, may require special equipment, hardware, or software or stronger internet access) [25]

However, patients and clinicians reported insufficient technical support, infrastructure, or equipment to access telehealth and difficulty with more complex telehealth systems that required special hardware or software support [25,32]. Some patients reported difficulty finding privacy at home to attend teleconsultations [21]. Some patients felt excluded or deterred from seeking help because of unfamiliarity with technology [21,22]. Some clinicians reported drawbacks of telehealth, such as it being harder to address language or cognition barriers with patients without physical cues [22] as well as feeling concerned with risks on digital platforms (eg, cyberattacks, security, and confidentiality in web-based communication) [25].

Effectiveness

The NQF outcome measure “Effectiveness” (ie, represents the systematic, clinical, operational, and technical success or barriers of telehealth) was reported in 84% (16/19) of the included studies. A summary of benefits and drawbacks of telehealth per this outcome factor is provided in Table 6. Both patients and clinicians reported that telehealth was suitable (ie, clinical appropriateness) for infections, dermatological concerns, renewal of prescriptions, or self-monitoring programs [25,32,33]. Most patients reported being sufficient at self-assessing whether they should seek a teleconsultation or an in-person consultation according to their health concerns [25]. Furthermore, patients could show their medication or self-care practices at home, allowing clinicians to better understand how their home environment may affect their self-management, thus improving clinicians’ advice dispensed to support their patients [1]. Clinicians also noted the benefits of sharing medical records with patients via screen sharing, improving their understanding [19].

However, telehealth was reported as not suitable for specific patient groups (eg, people with unstable mental concerns or low hearing and vision, young children unable to describe symptoms themselves, and people with cognitive impairment) [30] or for complex symptom presentations or diagnoses that required physical examinations (eg, chest pain, stomach pain, and potential new cancer) [1,9]. There is currently a lack of guidance on identifying and addressing severe adverse events that may occur because of telehealth (eg, lack of guidance on safety netting for teleconsultation, uncertainty about who else is also present but hiding during the teleconsultation, or recording consultations without consent) [1]. Furthermore, there is a tendency to rely more on patient-reported outcomes and patient-directed examinations during telehealth, affecting a GP’s assessment of the patient’s health status, which may inevitably result in in-person consultations later on despite having had a teleconsultation [25,33].

Table 6. Benefits and drawbacks of telehealth according to the “Effectiveness” outcome factor per perspective.

Perspective and benefits of telehealth	Drawbacks of telehealth
Primary care clinician perspective	
• Easier to share medical records with patients via screen sharing during video consultations [19] • More efficient consultations with patients (ie, focused discussions and pretriaging procedures to preidentify concerns) [32]	• Lack of guidance on appropriate ways to address serious adverse events related to telehealth [1] • Increased reliance on trusting patients’ reported symptoms and self-examination assessment [33] • Not suited for complex symptom presentations or diagnoses that require physical examinations (eg, chest pain, stomach pain, and potential new cancer) [1,9]
Patient perspective	
• Improved ability for patients to self-manage their health because of their ability to share their medications or self-care practices at home with their clinicians [1] • Most patients can self-assess the suitability of telehealth according to their health concerns [25]	• Still requiring in-person consultations despite having had a teleconsultation already [25]
Both primary care clinician and patient perspective	
• Perceived to be suitable for dermatological concerns and renewal of prescriptions or self-monitoring programs for improved patient outcomes [23,25]	• Unsuitable for certain at-risk patient groups (eg, people who are mentally unstable or have low hearing and vision, young children, and people with cognitive impairment) [31,33]

Experience

The NQF outcome measure “Experience” (ie, represents the usability and effect of telehealth on patients and providers) was reported in 89% (17/19) of the included studies. A summary of benefits and drawbacks of telehealth per this outcome factor is provided in Table 7.

Both clinicians and patients were satisfied with a perceived lower risk of infection transmission during the COVID-19 pandemic as a result of using telehealth [1,25-27,30,33]. Some reported feeling positive that they were able to maintain a patient-provider connection via telehealth during the COVID-19 pandemic [9,32,33]. Primary care clinicians reported several personal benefits of telehealth, including improved work-life balance and the ability to conduct some consultations more efficiently [33]. Clinicians also reported perceiving their patients as feeling more relaxed in their home environments compared with in-person consultations [33]. Overall, most patients reported having a satisfactory experience and a willingness to use telehealth again [21,24].

Table 7. Benefits and drawbacks of telehealth according to the “Experience” outcome factor per perspective.

Perspective and benefits of telehealth	Drawbacks of telehealth
Primary care clinician perspective	
• Improved work-life balance [33] • Satisfied in perceiving their patients to be more relaxed in telehealth settings [32] • Easier to conduct some consultations more efficiently [28]	• Concerned about cultural and language barriers with patients [23] • Lacking stimulating work for some clinicians as there is little in-person interaction with patients [9] • Reliant on clinicians taking on multiple roles (eg, secretary, IT support, and clinician) [26,33]
Patient perspective	
• Satisfactory experience with telehealth consultations for surveyed patients [24,31] • Surveyed patients willing to use telehealth again [21]	• Lacking opportunity to develop in-person rapport because of cultural or language barriers, technological barriers, and confidentiality concerns [25,29] • Lacking in establishing new patient-provider relationships [9,23] • Impersonal in comparison with in-person care because of the remote nature of telehealth [27,31]
Both primary care clinician and patient perspective	
• Satisfied with lower risk of infection transmission [1,25-27,30] • Positive patient-provider relationship for some patients as the personal connection was felt in teleconsultations [32]	• Dissatisfied with the lack of in-person physical examinations [9,24,33]

As reported by both clinicians and patients, the main drawback of telehealth was dissatisfaction with a lack of in-person physical examinations [24]. Some clinicians faced the additional drawbacks of addressing language or cultural barriers without in-person cues [23] and the lack of stimulating work when there was little in-person interaction with patients [9]. In addition, telehealth sometimes required clinicians to take on multiple roles in the practice to ensure it ran smoothly (eg, secretary, IT support, and clinician) [26,33]. Patients similarly needed to combat barriers such as the lack of opportunity to develop rapport with their clinicians, impersonal consultations [27], language or cultural barriers to disclosing issues, technological barriers, and confidential concerns during web-based communication via telehealth [23].

Financial Impact or Cost

The NQF outcome measure "Financial Impact/Cost" (ie, potential cost savings or losses to patients, families, or providers) was reported in 63% (12/19) of the included studies. A summary of benefits and drawbacks of telehealth per this outcome factor is provided in Table 8. From the clinicians' perspective, the infrastructure, processes, and long-term reimbursement models of telehealth were important considerations before its full potential and benefits could be unleashed [9,28]. For patients, removing the need to travel and reducing the loss of pay from taking time off work to attend in-person consultations were important drivers for choosing telehealth [34].

Table 8. Benefits and drawbacks of telehealth according to the "Financial Impact/Cost" outcome factor per perspective.

Perspective and benefits of telehealth	Drawbacks of telehealth
Primary care clinician perspective	
• Reduced telehealth setup costs because of existing infrastructure and processes (eg, adequate funding model and absence of billing or licensure restrictions) [9,18] • Cost-effective in the long run because of reduced running costs compared with in-person consultations [9,22,31] • Reimbursement model available for teleconsultations (eg, Medicare support in Australia) [28,29]	• Expensive to set up a telehealth system from scratch [25] • Long-term funding models are not globally determined, potentially opening up opportunities for commercial entities to exploit [25]
Patient perspective	
• Some patients prefer telehealth consultations and are willing to pay [34] • Some patients report that telehealth consultation fees are appropriate [24,31] • Saving costs using telehealth (eg, travel costs to in-person clinics and for patients needing to take time off work for appointments) [34]	• Mixed responses from some patients regarding willingness to pay for teleconsultation [26] • Inappropriate telehealth consultation charges felt by some patients [24]

However, issues relating to long-term models of financing and reimbursing telehealth remained a major concern [25]. For both patients and clinicians, there were concerns that remain to be researched about the expensive costs of acquiring the necessary software, hardware, and infrastructure to set up telehealth when it is unclear whether telehealth will remain a permanent service delivery mode in the long term. Furthermore, there is potential for commercial entities to exploit the charging or provision of telehealth when there remains uncertainty from the government on its long-term funding model [25]. There were mixed views regarding whether patients were willing to pay the same rate for telehealth consultations when compared with in-person consultations or an alternative appropriate cost [31].

Discussion

Principal Findings

To our knowledge, this is the first systematic review reporting visit types in primary care where telehealth was used during the COVID-19 pandemic. Most of the included studies (13/19, 68%) were level-4 evidence (cohort studies, interviews, and surveys), reflecting the early experience of the pandemic. Seven primary care visit types were identified: chronic condition management (17/19, 89%), existing patients (17/19, 89%), medication management (17/19, 89%), new patients (16/19, 84%), mental health and behavioral management (15/19, 79%), post–test result follow-up (14/19, 74%), and postdischarge follow-up (7/19, 37%). Benefits and drawbacks were reported across all visit types, with chronic condition management visits being one of the visit types with use of telehealth reporting the greatest number of studies during the pandemic (17/19, 89%). Reasons for why telehealth was deemed suitable for chronic condition management visits included patients having pre-existing diagnoses, established patient-provider relationships, and lack of complex physical examinations required. Insights into both the primary care clinician and patient perspective of telehealth use for specific visit types (ie, access to care, effectiveness, experience, and financial impact or cost) were also provided. Overall, benefits of telehealth included improved convenience, focused discussions, and continuity of care despite social distancing practices during the COVID-19 pandemic. Drawbacks of telehealth included technical barriers, impersonal interactions, and semi-established reimbursement models.

Strengths and Limitations

The strengths of this study include following a rigorous approach at all stages of the systematic review. For example, a wide range of academic databases and gray literature were searched to ensure great coverage of literature. In total, 3 independent researchers following predetermined eligibility criteria were involved in article screening to reduce the risks of selection bias. Data extraction templates (eg, the Joanna Briggs Institute data abstraction form) were used to standardize reporting of findings between studies. Well-established tools (eg, GRADE and the Mixed Methods Appraisal Tool) were used to conduct a critical appraisal and assess levels of evidence for each included study. Furthermore, definitions and terminologies from widely accepted frameworks in the telehealth and primary care communities (such as the NQF Telehealth Framework, the RACGP, and the Medicare Benefits Schedule) were used to ease the translation of our review.

The limitations of this review include restricting it to studies between late 2019 and August 2022 as definitions of the COVID-19 era, limiting it to studies written in English, and the decision to focus on broadness rather than narrowness in our search strategy. Publication bias (ie, the tendency to report positive results) may be present in the included studies because of the novel adoption of telehealth during the COVID-19 pandemic and growing interest in this research space [44]. Since our review, additional studies may have been published focusing on the experience of telehealth as GPs and patients have become more experienced with its use within routine primary care settings. Thus, despite multiple search cycles, our review may only reflect early experiences of telehealth during the COVID-19 era. In addition, this systematic review focuses on the early experience of the COVID-19 pandemic, where primary studies on clinical outcomes of using telehealth during the pandemic were not yet available. Future reviews should examine the long-term clinical outcomes of patients using telehealth (or hybrid models of care) in primary care settings. Our search strategy did not use keywords related to specific visit types in primary care. Instead, we chose to focus broadly on primary care to ensure we captured all studies with telehealth support conducted in primary care during the pandemic that were reported. Future reviews could include non-English studies or specific visit types to increase the generalizability and scope of the findings.

Comparison With Prior Work

Before the COVID-19 pandemic, studies on telehealth focused on issues such as particular visit types (eg, medication reviews or chronic condition management visits) [45,49], patient satisfaction [46,47], or nonsynchronous patient-provider communication (eg, e-consultation portals) [48]. For example, a review by Polisina et al [45] explored the use of an at-home management program for a chronic condition such as diabetes. A systematic review by Hanjani et al [49] focused explicitly on medication reviews via telehealth and identified similar facilitators and barriers to those of our review. Most of the benefits and drawbacks of telehealth reported in this review, such as ease of use, reduced travel times, low cost, and improved communication (in some instances), were also found by Kruse et al [46] in their systematic review. Other studies such as that by Hollander and Goldwater [47] examined the use of telehealth in orthopedic surgery, and Villarreal et al [48] reviewed mobile systems designed for health care monitoring.

Our systematic review focused on studies published in the COVID-19 era to consider how telehealth was used in primary care during the pandemic. A recent systematic review by Snoswell et al [15] aligns with our recommendations, stating that telehealth services are equivalent to or (at times) more effective than in-person care. However, Snoswell et al [15] did not report telehealth experience during the COVID-19 pandemic, instead focusing on studies from 2010 to 2019. A recent systematic review from Carrillo de Albornoz et al [50] evaluated the effectiveness of teleconsultations in primary care and mental health services in comparison with in-person visits, providing similar insights into the usability of telehealth as an effective alternative to in-person consultations. However, although this study was published following the emergence of COVID-19, the included studies were not conducted during the COVID-19 pandemic and, therefore, this study does not reflect on the effectiveness of teleconsultations in light of the pandemic.

A rapid scoping review by Jonnagaddala et al [51] explored facilitators and inhibitors of primary care informatics to COVID-19 in Australia. Similarly, we found limited high-quality evidence on the effectiveness, access, equity, utility, safety, and quality of digital health during the COVID-19 pandemic. However, our review differs in the systematic review approach. We identified 7 visit types where telehealth was used in primary care during the pandemic, outlining the benefits and drawbacks of using telehealth for each visit type and in primary care overall.

Implications for Digital Health, Clinical Practice, and Future Research

In total, 3 key insights have emerged from this review.

Key Insight 1: Rigorous Research Is Needed to Investigate Which Visit Types Are Indeed Suitable for Telehealth in Primary care

The results of our systematic review identified a lack of quality evidence on primary care visit types suitable for telehealth. Most of the included studies (13/19, 68%) were level-4 evidence (ie, case series or cross-sectional studies), which are subject to self-report bias. Furthermore, there is a lack of focus on how telehealth was used for different visit types in primary care. The saturation of level-4 evidence in this space conveys that these study designs are indeed the current state of the art, presumably from the relatively short time since the start of the pandemic as well as the lack of ability to conduct follow-up or person-facing studies because of social distancing restrictions. As we move into the era of living with COVID-19, studies with a longitudinal follow-up that focus on specific visit types are required to assess the long-term suitability of telehealth in primary care. In addition, there was a lack of research in the included studies reporting clinical outcomes related to telehealth use during the pandemic, presumably because, at the time of searching and writing (ie, early phases of the COVID-19 pandemic), studies assessing clinical outcomes of using telehealth during the pandemic would not have yet been available. When those studies become available, future systematic reviews may wish to assess clinical outcomes of using telehealth during the pandemic so that the findings in this review reporting the suitability of telehealth for primary care visits can be validated.

Key Insight 2: Long-term Models of Telehealth and Their Impact on Patient Outcomes and Health Service Use

As a result of COVID-19, several countries (eg, Australia, the United States, and the United Kingdom) have introduced permanent or long-term funding for telehealth in primary care. For example, the Australian government introduced long-term funding for telehealth in December 2021 to align with initiatives to reduce community COVID-19 transmission [52]. Australians have welcomed telehealth consultations, with >86 million primary care telehealth consultations completed in Australia since the beginning of the COVID-19 pandemic [8]. Other countries such as the United States and the United Kingdom are exploring a permanent funding scheme for telehealth within their existing health care models. Almost every state Medicaid program has a reimbursement coverage account for telehealth services in the United States [53]. The Centers for Disease Control and Prevention also introduced multiple waivers during the COVID-19 pandemic to grant payment parity for telehealth [54].

Before the pandemic, telehealth policies in the United Kingdom alone were underdetermined across England, Wales, Northern Ireland, and Scotland. Challenges related to outdated systems and underinvestment in telehealth have hindered the progress of digitization [55]. During the COVID-19 pandemic, health care services under the National Health Service took a “total triage” approach where all patients were referred first to telehealth services over face-to-face services [55]. According to the Health Foundation, this initiative has caused a rapid and significant increase in telehealth use, reporting the highest-ever number of telephone consultations in English primary care as a consequence of the pandemic [56]. For example, a videoconferencing telehealth platform called “Near Me” reported having been used by approximately 300 people per week at the start of 2020, rising to approximately 20,000 appointments every week by mid-2020 [55]. By July 2021, >1 million appointments were delivered via telehealth services [55]. Furthermore, 11.4 million telephone consultations were reported to have been completed in March 2021 compared with 3.5 million in March 2019 [56]. This rapid and unforeseen uptake of telehealth services raises questions as to whether unintended consequences and safety risks may have been introduced as well [55].

Governments have recognized the value of telehealth during the pandemic, especially for patients who struggle with mobility [1,25], live remotely or rurally [15], or are unable to find suitable times to attend in-person consultations [8], regardless of their COVID-19 status. Future research ought to examine how long-term funding models of telehealth affect patient outcomes, help-seeking behaviors, and health service use patterns. For example, further research is required to compare the health outcomes and quality of care between patients who primarily use telehealth experiences versus those who use in-person care. In addition, further research is required to analyze how changes in health service use patterns because of routine telehealth use affect the funneling of resources, particularly training opportunities for health care providers on how to use telehealth optimally and the communication skills required in telehealth.

Furthermore, there is the additional consideration of how designs of telehealth need to evolve with emerging safety, ethical, and equitable concerns, for example, how to ensure that all patients can equally access care, regardless of the digital divide, if more resources are directed to providing telehealth over in-person services. In addition, further research is required to explore how to support patient-provider relationships when care is delivered across a blended model of approaches, as well as further research into appropriate safety-netting practices during teleconsultations [57].

Key Insight 3: Patient Safety at Home Is Paramount as Care and Technology Are Increasingly Used Outside Clinical Settings

Increasingly, care is moved closer to patients’ homes, blended with technology. The pandemic has accelerated the movement of blending care and technology at home. For example, home oximetry monitoring programs have been introduced for monitoring positive COVID-19 patients in the United Kingdom, and a recent prospective study has reported patient satisfaction and early success [58]. Other digital health services have also been increasingly introduced for use outside clinical settings, such as assistive technology to support independent living at home [13,59], remote monitoring mobile apps [60,61], and e–mental health services (eg, Betterhelp and Headspace) [52].

Introducing technologies directly into patients’ homes as part of routine service delivery may encourage more frequent monitoring of signs and symptoms. However, patient-facing medical devices and at-home care can introduce a new dimension of patient risk [62]. Historically, the role of conducting physical examinations and use of medical devices was reserved for health care professionals [63]. However, with telehealth and remote care services, the responsibility of physical examination and monitoring falls onto the patient, requiring patients to have the necessary knowledge and skills to conduct these previously clinician-directed tasks effectively by themselves or be aware of when to seek additional assistance [63]. As a result, patients could become vulnerable to unanticipated risks such as inaccurate self-examination [64,65], unreliable patient self-reports, reduced person-centered care because of language or cognitive barriers, inability to conduct a physical examination properly, or incapacity to receive care properly because of technological limitations [62,63]. Further investigation is required to identify the types of adverse events that can occur during remote care (eg, whether people are using technologies as intended or whether technologies are introducing unintended consequences) and ways to combat these adverse events [64].

Conclusions

This systematic review identified 7 visit types in primary care with telehealth support during the COVID-19 pandemic, with the greatest number of studies reporting benefit findings for *chronic condition management* visits (17/19, 89%). Benefits and drawbacks of using telehealth were reported across different visit types from patient and clinician perspectives, as well as the circumstances in which telehealth was found to be suitable (or not) for each primary care visit type. As telehealth potentially becomes a long-term care delivery model, improving telehealth consultation delivery while monitoring patient safety at home will emerge as an important priority area.

Acknowledgments

KW was supported with a Higher Degree Research scholarship from Macquarie University. AYSL was supported by the National Health and Medical Research Council grant APP1134919 (Centre for Research Excellence in Digital Health) and the New South Wales Health Early-Mid Career Fellowship.

Authors' Contributions

KW and AYSL contributed to the conception of the study. KW, FS, and NNK were involved in the literature search, screening, and data extraction. KW conducted the analyses and manuscript write-up with support from AYSL and SV. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 26 KB - medinform_v10i11e40469_app1.docx]

Multimedia Appendix 2
Study Characteristics.
[DOCX File , 51 KB - medinform_v10i11e40469_app2.docx]

Multimedia Appendix 3
Search Strategy.
[DOCX File , 620 KB - medinform_v10i11e40469_app3.docx]

Multimedia Appendix 4
Eligibility Criteria.
[DOCX File , 15 KB - medinform_v10i11e40469_app4.docx]

Multimedia Appendix 5
Data Extraction Template.
[DOCX File , 16 KB - medinform_v10i11e40469_app5.docx]

Multimedia Appendix 6
Risk and Bias Critical Appraisal.
[DOCX File , 26 KB - medinform_v10i11e40469_app6.docx]

Multimedia Appendix 7
Visit Type Medicare Descriptions.
[DOCX File , 32 KB - medinform_v10i11e40469_app7.docx]

Multimedia Appendix 8
Supporting Evidence of Suitability of Telehealth for each Visit Type during COVID-19.
[DOCX File , 55 KB - medinform_v10i11e40469_app8.docx]

Multimedia Appendix 9
Supporting Evidence of Benefit and Drawback of using Telehealth in Primary care during COVID-19 according to outcomes of the NQF Framework.
[DOCX File , 28 KB - medinform_v10i11e40469_app9.docx]

References

1. Johnsen TM, Norberg BL, Kristiansen E, Zanaboni P, Austad B, Krogh FH, et al. Suitability of video consultations during the COVID-19 pandemic lockdown: cross-sectional survey among Norwegian general practitioners. J Med Internet Res 2021 Feb 08;23(2):e26433 [FREE Full text] [doi: 10.2196/26433] [Medline: 33465037]

2. Bavli I, Sutton B, Galea S. Harms of public health interventions against covid-19 must not be ignored. BMJ 2020 Nov 02;371:m4074. [doi: 10.1136/bmj.m4074] [Medline: 33139247]
3. Douglas M, Katikireddi SV, Taulbut M, McKee M, McCartney G. Mitigating the wider health effects of covid-19 pandemic response. BMJ 2020 Apr 27;369:m1557 [FREE Full text] [doi: 10.1136/bmj.m1557] [Medline: 32341002]
4. Nickelson DW. Telehealth and the evolving health care system: strategic opportunities for professional psychology. Prof Psychol Res Pract 1998 Dec;29(6):527-535. [doi: 10.1037/0735-7028.29.6.527]
5. Snoswell CL, Caffery LJ, Haydon HM, Thomas EE, Smith AC. Telehealth uptake in general practice as a result of the coronavirus (COVID-19) pandemic. Aust Health Rev 2020 Sep;44(5):737-740. [doi: 10.1071/AH20183] [Medline: 32853536]
6. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res 2020 Dec 01;22(12):e24087 [FREE Full text] [doi: 10.2196/24087] [Medline: 33147166]
7. Tsirtsakis A. GPs have embraced telehealth, survey finds. The Royal Australian College of General Practitioners. 2020 Aug 6. URL: https://www1.racgp.org.au/newsgp/professional/gps-have-embraced-telehealth-survey-finds#:~:text=COVID%2D19%20virus.-,',seek%20the%20care%20they%20need [accessed 2021-02-14]
8. Evans J. Telehealth services to become a permanent part of healthcare system, following COVID-19 success. ABC News. 2021 Dec 13. URL: https://www.abc.net.au/news/2021-12-13/telehealth-services-to-be-made-permanent/100694844 [accessed 2022-01-25]
9. De Guzman KR, Snoswell CL, Giles CM, Smith AC, Haydon HH. GP perceptions of telehealth services in Australia: a qualitative study. BJGP Open 2022 Mar;6(1):BJGPO.2021.0182 [FREE Full text] [doi: 10.3399/BJGPO.2021.0182] [Medline: 34819294]
10. Reeves JJ, Ayers JW, Longhurst CA. Telehealth in the COVID-19 era: a balancing act to avoid harm. J Med Internet Res 2021 Feb 01;23(2):e24785 [FREE Full text] [doi: 10.2196/24785] [Medline: 33477104]
11. Duckett S. What should primary care look like after the COVID-19 pandemic? Aust J Prim Health 2020 Jun;26(3):207-211. [doi: 10.1071/PY20095] [Medline: 32454003]
12. Mroz G, Papoutsi C, Rushforth A, Greenhalgh T. Changing media depictions of remote consulting in COVID-19: analysis of UK newspapers. Br J Gen Pract 2020 Dec 28;71(702):e1-e9 [FREE Full text] [doi: 10.3399/BJGP.2020.0967] [Medline: 33318086]
13. Donaghy E, Atherton H, Hammersley V, McNeilly H, Bikker A, Robbins L, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019 Sep;69(686):e586-e594 [FREE Full text] [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
14. Shah AC, Badawy SM. Telemedicine in pediatrics: systematic review of randomized controlled trials. JMIR Pediatr Parent 2021 Feb 24;4(1):e22696 [FREE Full text] [doi: 10.2196/22696] [Medline: 33556030]
15. Snoswell CL, Chelberg G, De Guzman KR, Haydon HH, Thomas EE, Caffery LJ, et al. The clinical effectiveness of telehealth: a systematic review of meta-analyses from 2010 to 2019. J Telemed Telecare 2021 Jun 29:1357633X211022907. [doi: 10.1177/1357633X211022907] [Medline: 34184580]
16. Morgan RL, Whaley P, Thayer KA, Schünemann HJ. Identifying the PECO: a framework for formulating good questions to explore the association of environmental and other exposures with health outcomes. Environ Int 2018 Dec;121(Pt 1):1027-1031 [FREE Full text] [doi: 10.1016/j.envint.2018.07.015] [Medline: 30166065]
17. Hollander J, Ward M, Alverson D, Bashshur R, Darkins A, DePhillips H. Creating a framework to support measure development for Telehealth. National Quality Forum. Washington, DC, USA: National Quality Forum; 2017 Aug 31. URL: https://www.aristamd.com/wp-content/uploads/2018/10/telehealth_final_report.pdf [accessed 2021-03-20]
18. Jetty A, Jabbarpour Y, Westfall M, Kamerow DB, Petterson S, Westfall JM. Capacity of primary care to deliver telehealth in the United States. J Am Board Fam Med 2021 Feb;34(Suppl):S48-S54 [FREE Full text] [doi: 10.3122/jabfm.2021.S1.200202] [Medline: 33622818]
19. Grossman Z, Chodick G, Reingold SM, Chapnick G, Ashkenazi S. The future of telemedicine visits after COVID-19: perceptions of primary care pediatricians. Isr J Health Policy Res 2020 Oct 20;9(1):53 [FREE Full text] [doi: 10.1186/s13584-020-00414-0] [Medline: 33081834]
20. Jabbarpour Y, Jetty A, Westfall M, Westfall J. Not telehealth: which primary care visits need in-person care? J Am Board Fam Med 2021 Feb;34(Suppl):S162-S169 [FREE Full text] [doi: 10.3122/jabfm.2021.S1.200247] [Medline: 33622832]
21. van de Poll-Franse LV, de Rooij BH, Horevoorts NJ, May AM, Vink GR, Koopman M, et al. Perceived care and well-being of patients with cancer and matched norm participants in the COVID-19 crisis: results of a survey of participants in the Dutch PROFILES registry. JAMA Oncol 2021 Feb 01;7(2):279-284 [FREE Full text] [doi: 10.1001/jamaoncol.2020.6093] [Medline: 33237294]
22. Gomez T, Anaya YB, Shih KJ, Tarn DM. A qualitative study of primary care physicians' experiences with telemedicine during COVID-19. J Am Board Fam Med 2021 Feb;34(Suppl):S61-S70 [FREE Full text] [doi: 10.3122/jabfm.2021.S1.200517] [Medline: 33622820]
23. Hasani SA, Ghafri TA, Al Lawati H, Mohammed J, Al Mukhainai A, Al Ajmi F, et al. The use of telephone consultation in primary health care during COVID-19 pandemic, Oman: perceptions from physicians. J Prim Care Community Health 2020;11:2150132720976480 [FREE Full text] [doi: 10.1177/2150132720976480] [Medline: 33307943]

24. Imlach F, McKinlay E, Middleton L, Kennedy J, Pledger M, Russell L, et al. Telehealth consultations in general practice during a pandemic lockdown: survey and interviews on patient experiences and preferences. BMC Fam Pract 2020 Dec 13;21(1):269 [FREE Full text] [doi: 10.1186/s12875-020-01336-1] [Medline: 33308161]
25. Gabrielsson-Järhult F, Kjellström S, Josefsson KA. Telemedicine consultations with physicians in Swedish primary care: a mixed methods study of users' experiences and care patterns. Scand J Prim Health Care 2021 Jun;39(2):204-213 [FREE Full text] [doi: 10.1080/02813432.2021.1913904] [Medline: 33974502]
26. Murphy M, Scott LJ, Salisbury C, Turner A, Scott A, Denholm R, et al. Implementation of remote consulting in UK primary care following the COVID-19 pandemic: a mixed-methods longitudinal study. Br J Gen Pract 2021 Feb 25;71(704):e166-e177 [FREE Full text] [doi: 10.3399/BJGP.2020.0948] [Medline: 33558332]
27. Schweiberger K, Hoberman A, Iagnemma J, Schoemer P, Squire J, Taormina J, et al. Practice-level variation in telemedicine use in a pediatric primary care network during the COVID-19 pandemic: retrospective analysis and survey study. J Med Internet Res 2020 Dec 18;22(12):e24345 [FREE Full text] [doi: 10.2196/24345] [Medline: 33290244]
28. New items for COVID-19 telehealth and phone services. The Royal Australian College of General Practitioners. 2021 May 17. URL: https://www.racgp.org.au/running-a-practice/practice-resources/medicare/medicare-benefits-schedule/new-items-for-covid-19-telehealth-services [accessed 2021-03-03]
29. MBS Changes factsheet - COVID-19 temporary MBS telehealth services. Department of Health, Australian Government. 2020 Sep 18. URL: http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/0C514FB8C9FBBEC7CA25852E00223AFE/$File/Factsheet-COVID-19-Bulk-billed-MBS%20telehealth-Services-Overarching-17.09.2020.pdf [accessed 2021-03-03]
30. Mozes I, Mossinson D, Schilder H, Dvir D, Baron-Epel O, Heymann A. Patients' preferences for telemedicine versus in-clinic consultation in primary care during the COVID-19 pandemic. BMC Prim Care 2022 Feb 22;23(1):33 [FREE Full text] [doi: 10.1186/s12875-022-01640-y] [Medline: 35193509]
31. Javanparast S, Roeger L, Kwok Y, Reed RL. The experience of Australian general practice patients at high risk of poor health outcomes with telehealth during the COVID-19 pandemic: a qualitative study. BMC Fam Pract 2021 Apr 08;22(1):69 [FREE Full text] [doi: 10.1186/s12875-021-01408-w] [Medline: 33832422]
32. Assing Hvidt E, Christensen NP, Grønning A, Jepsen C, Lüchau EC. What are patients' first-time experiences with video consulting? A qualitative interview study in Danish general practice in times of COVID-19. BMJ Open 2022 Apr 15;12(4):e054415 [FREE Full text] [doi: 10.1136/bmjopen-2021-054415] [Medline: 35428624]
33. Due TD, Thorsen T, Andersen JH. Use of alternative consultation forms in Danish general practice in the initial phase of the COVID-19 pandemic - a qualitative study. BMC Fam Pract 2021 Jun 02;22(1):108 [FREE Full text] [doi: 10.1186/s12875-021-01468-y] [Medline: 34078281]
34. Manski-Nankervis JA, Davidson S, Hiscock H, Hallinan C, Ride J, Lingam V, et al. Primary care consumers’ experiences and opinions of a telehealth consultation delivered via video during the COVID-19 pandemic. Aust J Prim Health 2022 Jun;28(3):224-231. [doi: 10.1071/PY21193] [Medline: 35287793]
35. The EndNote Team. EndNote. Clarivate. Philadelphia, PA, USA: Clarivate; 2013. URL: https://endnote.com/ [accessed 2021-03-25]
36. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
37. Peters MD, Godfrey C, McInerney P, Mun Z, Tricco AC, Khalil H. Appendix 11.1 JBI template source of evidence details, characteristics and results extraction instrument. In: Aromataris E, Munn Z, editors. Joanna Briggs Institute Reviewer's Manual. Adelaide, Australia: Joanna Briggs Institute; 2017.
38. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
39. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011 Apr;64(4):380-382. [doi: 10.1016/j.jclinepi.2010.09.011] [Medline: 21185693]
40. Jiang M, Xie YQ, Xie JX, Zou XW, Lai KF. Methodology for development of the Chinese evidence-based clinical practice guideline of the diagnosis and management of cough. J Thorac Dis 2018 Nov;10(11):6310-6313 [FREE Full text] [doi: 10.21037/jtd.2018.09.154] [Medline: 30623940]
41. Medicare benefits schedule book operating from January 1 2020. Department of Health, Australian Government. 2021 Jul 21. URL: http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/8F3FA58ED97DCA35CA2584BE00111151/$File/202001-MBS%2017Jan2020.pdf [accessed 2021-07-23]
42. COVID-19 Telehealth MBS items. MBS Online. Phillip, Australia: Department of Health and Ageing, Australian Government; 2020 Aug 13. URL: http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/news-2020-03-29-latest-news-March [accessed 2021-07-23]
43. Chronic Disease Management - Provider Information. The Department of Health and Aged Care. Phillip, Australia: Department of Health and Aged Care, Australian Government; 2016 Sep 2. URL: https://www1.health.gov.au/internet/main/publishing.nsf/Content/mbsprimarycare-factsheet-chronicdisease.htm [accessed 2021-07-23]

44. DeVito NJ, Goldacre B. Catalogue of bias: publication bias. BMJ Evid Based Med 2019 Apr;24(2):53-54. [doi: 10.1136/bmjebm-2018-111107] [Medline: 30523135]
45. Polisena J, Coyle D, Coyle K, McGill S. Home telehealth for chronic disease management: a systematic review and an analysis of economic evaluations. Int J Technol Assess Health Care 2009 Jul;25(3):339-349. [doi: 10.1017/S0266462309990201] [Medline: 19619353]
46. Kruse CS, Krowski N, Rodriguez B, Tran L, Vela J, Brooks M. Telehealth and patient satisfaction: a systematic review and narrative analysis. BMJ Open 2017 Aug 03;7(8):e016242 [FREE Full text] [doi: 10.1136/bmjopen-2017-016242] [Medline: 28775188]
47. Hollander JE, Goldwater J. Telemedicine research and quality assessment. In: Atanda Jr A, Lovejoy III JF, editors. Telemedicine in Orthopedic Surgery and Sports Medicine: Development and Implementation in Practice. Cham, Switzerland: Springer; 2021:157-168.
48. Villarreal V, Hervás R, Bravo J. A systematic review for mobile monitoring solutions in m-health. J Med Syst 2016 Sep;40(9):199. [doi: 10.1007/s10916-016-0559-5] [Medline: 27464519]
49. Shafiee Hanjani L, Caffery LJ, Freeman CR, Peeters G, Peel NM. A scoping review of the use and impact of telehealth medication reviews. Res Social Adm Pharm 2020 Aug;16(8):1140-1153. [doi: 10.1016/j.sapharm.2019.12.014] [Medline: 31874815]
50. Carrillo de Albornoz S, Sia KL, Harris A. The effectiveness of teleconsultations in primary care: systematic review. Fam Pract 2022 Jan 19;39(1):168-182 [FREE Full text] [doi: 10.1093/fampra/cmab077] [Medline: 34278421]
51. Jonnagaddala J, Godinho MA, Liaw ST. From telehealth to virtual primary care in Australia? A Rapid scoping review. Int J Med Inform 2021 Jul;151:104470. [doi: 10.1016/j.ijmedinf.2021.104470] [Medline: 34000481]
52. Scott A, Bai T, Zhang Y. Association between telehealth use and general practitioner characteristics during COVID-19: findings from a nationally representative survey of Australian doctors. BMJ Open 2021 Mar 24;11(3):e046857 [FREE Full text] [doi: 10.1136/bmjopen-2020-046857] [Medline: 33762248]
53. Fact Sheet: Telehealth. American Hospital Association. 2019 Feb 1. URL: https://www.aha.org/factsheet/telehealth [accessed 2022-02-01]
54. Using Telehealth to Expand Access to Essential Health Services during the COVID-19 Pandemic. Centers for Disease Control and Prevention. 2020 Jun 10. URL: https://www.cdc.gov/coronavirus/2019-ncov/hcp/telehealth.html [accessed 2022-02-01]
55. Hutchings R. The impact of Covid-19 on the use of digital technology in the NHS. Nuffield Trust. 2020 Aug 27. URL: https://www.nuffieldtrust.org.uk/research/the-impact-of-covid-19-on-the-use-of-digital-technology-in-the-nhs [accessed 2022-02-01]
56. Fraser C, Fisher R. How has the COVID-19 pandemic impacted primary care? The Health Foundation. 2021 May 27. URL: https://www.health.org.uk/news-and-comment/charts-and-infographics/how-has-the-covid-19-pandemic-impacted-primary-care [accessed 2022-09-05]
57. Greenhalgh T, Rosen R, Shaw SE, Byng R, Faulkner S, Finlay T, et al. Planning and evaluating remote consultation services: a new conceptual framework incorporating complexity and practical ethics. Front Digit Health 2021 Aug 13;3:726095 [FREE Full text] [doi: 10.3389/fdgth.2021.726095] [Medline: 34713199]
58. Clarke J, Flott K, Fernandez Crespo R, Ashrafian H, Fontana G, Benger J, et al. Assessing the safety of home oximetry for COVID-19: a multisite retrospective observational study. BMJ Open 2021 Sep 14;11(9):e049235 [FREE Full text] [doi: 10.1136/bmjopen-2021-049235] [Medline: 34521666]
59. Hospital in the home (HITH) - Performance. NSW Health. 2021 Aug 17. URL: https://www.health.nsw.gov.au/Performance/Pages/hith.aspx [accessed 2022-01-18]
60. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. NPJ Digit Med 2020 May 4;3:65 [FREE Full text] [doi: 10.1038/s41746-020-0280-0] [Medline: 32377576]
61. Clay-Williams R, Rapport F, Braithwaite J. The Australian health system response to COVID-19 from a resilient health care perspective: what have we learned? Public Health Res Pract 2020 Dec 09;30(4):3042025 [FREE Full text] [doi: 10.17061/phrp3042025] [Medline: 33294901]
62. Guise V, Anderson J, Wiig S. Patient safety risks associated with telecare: a systematic review and narrative synthesis of the literature. BMC Health Serv Res 2014 Nov 25;14:588 [FREE Full text] [doi: 10.1186/s12913-014-0588-z] [Medline: 25421823]
63. Stokke R, Melby L, Isaksen J, Obstfelder A, Andreassen H. A qualitative study of what care workers do to provide patient safety at home through telecare. BMC Health Serv Res 2021 Jun 05;21(1):553 [FREE Full text] [doi: 10.1186/s12913-021-06556-4] [Medline: 34090450]
64. Stevens WJ, van der Sande R, Beijer LJ, Gerritsen MG, Assendelft WJ. eHealth apps replacing or complementing health care contacts: scoping review on adverse effects. J Med Internet Res 2019 Mar 01;21(3):e10736 [FREE Full text] [doi: 10.2196/10736] [Medline: 30821690]
65. Akhtar M, Van Heukelom PG, Ahmed A, Tranter RD, White E, Shekem N, et al. Telemedicine physical examination utilizing a consumer device demonstrates poor concordance with in-person physical examination in emergency department

patients with sore throat: a prospective blinded study. Telemed J E Health 2018 Oct;24(10):790-796 [FREE Full text] [doi: 10.1089/tmj.2017.0240] [Medline: 29470127]

Abbreviations

GP: general practitioner
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
NQF: National Quality Forum
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RACGP: Royal Australian College of General Practitioners

Edited by C Lovis; submitted 22.06.22; peer-reviewed by E Donaghy, C Kruse, A Ritchie; comments to author 27.07.22; revised version received 16.08.22; accepted 02.10.22; published 28.11.22.

Please cite as:
Ward K, Vagholkar S, Sakur F, Khatri NN, Lau AYS
Visit Types in Primary Care With Telehealth Use During the COVID-19 Pandemic: Systematic Review
JMIR Med Inform 2022;10(11):e40469
URL: https://medinform.jmir.org/2022/11/e40469
doi:10.2196/40469
PMID:36265039

Viewpoint

Realizing the Potential of Computer-Assisted Surgery by Embedding Digital Twin Technology

Jiaxin Qin[1], ME; Jian Wu[1], PhD

Institute of Biomedical Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

Corresponding Author:
Jian Wu, PhD
Institute of Biomedical Engineering
Shenzhen International Graduate School
Tsinghua University
Building L, 3th Fl.
Taoyuan Street, University Town
Shenzhen, 518055
China
Phone: 86 13751113096
Email: wuj@sz.tsinghua.edu.cn

Abstract

The value of virtual world and digital phenotyping has been demonstrated in several fields, and their applications in the field of surgery are worthy of attention and exploration. This viewpoint describes the necessity and approach to understanding the deeper potential of computer-assisted surgery through interaction and symbiosis between virtual and real spaces. We propose to embed digital twin technology into all aspects of computer-assisted surgery rather than just the surgical object and further apply it to the whole process from patient treatment to recovery. A more personalized, precise, and predictable surgery is our vision.

(JMIR Med Inform 2022;10(11):e35138) doi:10.2196/35138

KEYWORDS

computer-assisted surgery; digital twin; virtual space; surgical navigation; remote surgery

Introduction

As is well known, computer-assisted surgery (CAS) has been widely used in the field of surgery. It enables the precise positioning and visualization of the patient's deconstructive structures and surgical instruments by integrating medical imaging, spatial positioning technology, and various computer technologies. Therefore, with CAS, the surgeon can plan the surgical procedure in a precise manner before the operation, such as the access point, implant selection, and placement [1]. However, intraoperative planning cannot be implemented accurately during surgery due to intraoperative disturbances such as respiratory movements and tissue deformation. Moreover, CAS is still lacking in postoperative prediction, prevention of surgical complications, and postoperative evaluation.

Digital twin (DT) is a concept or technology that refers to create a multiphysical, multiscale, and high-fidelity virtual representation of physical entity in the virtual space. We call this virtual representation the DT of its corresponding physical entity. The virtual representation can be dynamically updated when the physical entities change, enabling real-time mapping from physical space to virtual space. DT originates in the industry [2] and is now used in many fields such as precision health [3], manufacturing, construction, product design, and weather prediction [4]. Considering the features of CAS and DT, we propose to bring DT to CAS and explain why and how DT can be used to enhance the application of CAS in all its phases, especially the potential value in remote surgery when combining CAS with DT.

Bringing DT to CAS

Figure 1 illustrates our proposed DT-based CAS solution. We collect data related to surgical objects, surgical instruments, and medical devices in the preoperative period, and use these data to build their multiphysical, multiscale, and high-fidelity DT models in virtual space by mathematical simulation and modeling [5]. DT models can dynamically simulate a wide range of properties of the patient's tissues and organs, such as geometric, physical, physiological, and behavioral properties. Therefore, preoperative planning, surgical simulation, and postoperative prediction based on these models will be more accurate and comprehensive.

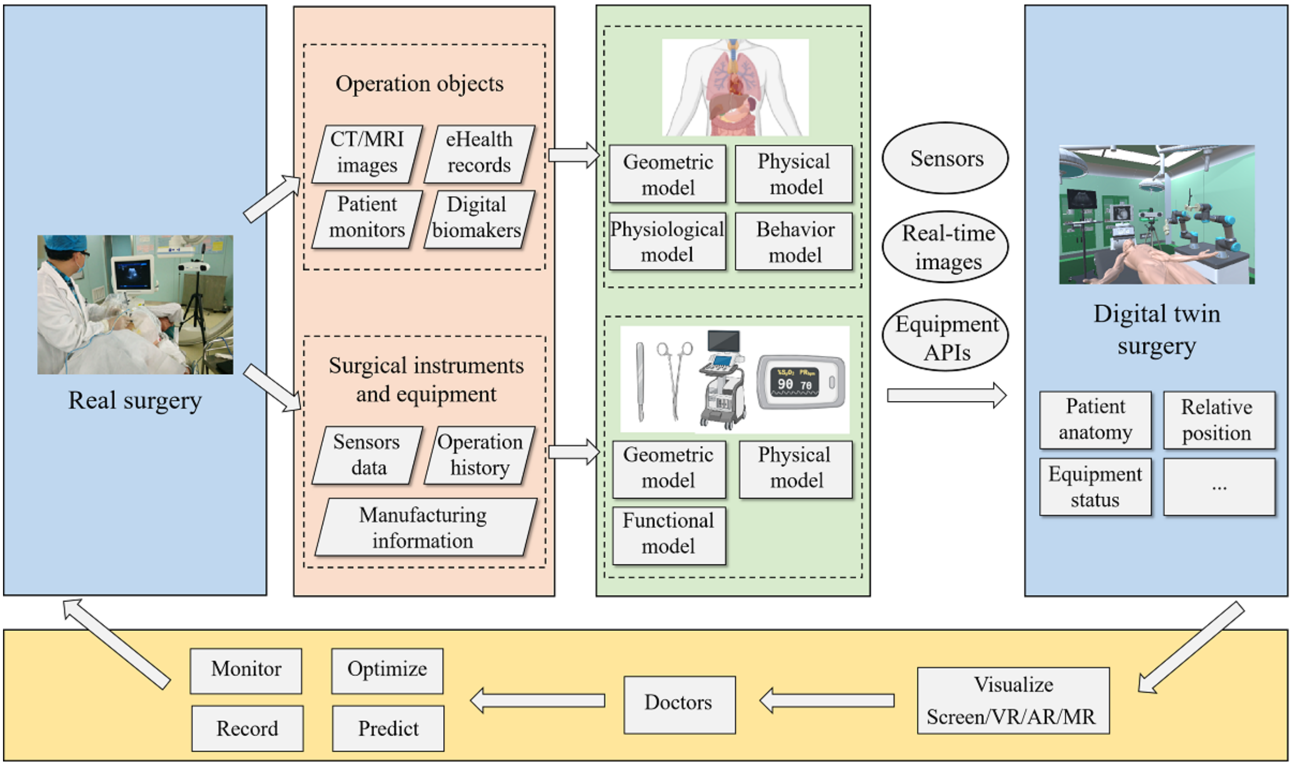
Information, as well as the changes of information from surgical instruments and the patient (such as anatomical structure, position, posture, etc), are acquired during surgery using real-time imaging technologies and a variety of sensors. Due to the high requirements for real-time performance, it is much more difficult to obtain whole information intraoperatively compared to preoperatively. For example, the clarity of the patient's anatomical structure obtained by ultrasound imaging is much less than that of preoperative CT (computed tomography) and MRI (magnetic resonance imaging). However, we can quickly capture local feature information from ultrasound images through advanced methods such as deep learning [6]. Combining this local information with the preoperative DT models, which already have physical, physiological, and behavioral characteristics, it is possible to obtain the overall dynamic changes of the patient's anatomical structure. In this way, intraoperative changes can be evaluated in real time, and all the information in the virtual space can be visualized when using extended reality (virtual reality, augmented reality, and mixed reality), and it also provides the surgeon with real-time, accurate surgical navigation.

Sensors can collect a variety of data that are needed to update the DT during the procedure. Commonly used sensors are physiological, mechanical, and position sensors. Among them, position sensors play a key role in positioning and tracking in surgical navigation and can help achieve a baseline position mapping relationship between the real surgical space and the DT surgical space. Optical sensors have higher accuracy and real-time performance compared with other position sensors. They are able to collect position and posture data of the optical marker in real time in the form of quaternions or transformation matrix. We usually fix the optical marker on the object of interest. For example, we can fix the optical marker on the ultrasound probe before the procedure and obtain the conversion between the optical marker coordinate system and the ultrasound image coordinate system by ultrasound probe calibration [7]. Intraoperatively, we can not only collect real-time position and posture information of the ultrasound probe in real time but also obtain the spatial position of any point displayed on the ultrasound image.

An application programming interface is a software intermediary that can create data links between different devices. We can use application programming interfaces of medical devices to obtain operational data in real-time and update the DTs. Through these DTs, we can monitor, control, and manage medical devices in a uniform way.

While the surgery is being performed in real space, a digital record of the surgery is updated simultaneously in virtual space, which we call the process twin of the whole surgery process. This record can be used for postoperative evaluation of the patient, including the evaluation of the actual surgery, the surgical procedure, the choice of strategy, and so on. It can also act as an important reference for postoperative follow-up.

Figure 1. Digital twin–based computer-assisted surgery (CAS) solution. With the support of digital twin surgery, we can perform monitoring, optimization, recording, and prediction for the real surgery. API: application programming interface; AR: augmented reality; CT: computed tomography; MR: mixed reality; MRI: magnetic resonance imaging; VR: virtual reality.



Potential in Remote Surgery

It is the establishment of an accurate match between the real world and the virtual space that enables effective remote surgery on real patients using DTs. The surgeon can obtain the real-time situation of the surgery at the remote end by the dynamically updated DTs and can use a haptic device to remotely control the robotic arm to perform the surgery [8] through high-speed, low-latency communication technologies such as 5G. Since the DTs contain rich information about the surgical object, the surgeon can obtain important feedback information [9] at the remote end through the haptic device without any other sensors. For each surgical device, it can also be monitored and controlled remotely through their DTs in virtual space.

Data Processing and Security

Data-driven approach is one of the core approaches to implement DT. The process of dynamically updating the DTs intraoperatively requires a large amount of data computation and interchange. By using cloud computing technology [10], hardware costs can be effectively reduced. Medical Cyber Physical Systems are the networked health care integration of medical devices [11]. They provide a superior way to capture, store, and securely access large amounts of medical data. In the future, its development may provide important data support for the application of DT [12]. Additional attention needs to be paid to the fact that the collection of private health data on human individuals may raise complex ethical issues [13]. Thus, data security should be carefully considered for the storage and retrieval of operation data, and data process encryption should be embodied in DT.

Discussion

Bringing DT to CAS is to monitor, optimize, record, and predict the surgical process in the real space by creating a virtual twin surgical space (including the DTs of the surgical object, surgical instruments, and medical equipment). In this way, it can leverage and integrate data from the entire surgical phase and is applied to the patient from treatment to recovery. This also determines the higher level of complexity of the DT systems. Different types of data from multiple devices need to be integrated within the same system and ensure the system's stable operation. The real-time dynamic response of the DTs requires high data transmission speed and network speed, especially in the remote surgery. In addition, the simulation of complex physiological signals of biological tissues is still a challenge that needs to be faced if a more detailed patient model is desired. In the future, the virtual twin space can be used as a carrier to establish an integrated, digital surgical process management system and to form a new clinical implementation system. Under such a system, surgical treatment will be more personalized, precise, and predictable.

Acknowledgments

The authors would like to thank the support of the Shenzhen Basic Research Project (JCY20200109142805928, JCYJ20160428182053361) and the National Key Research and Development Program of China (2019YFC0119500).

Conflicts of Interest

None declared.

References

1. Joskowicz L. Computer-aided surgery meets predictive, preventive, and personalized medicine. EPMA J 2017 Mar 07;8(1):1-4 [FREE Full text] [doi: 10.1007/s13167-017-0084-8] [Medline: 28670350]
2. Tao F, Zhang H, Liu A, Nee AYC. Digital twin in industry: State-of-the-art. IEEE Trans. Ind. Inf 2019 Apr;15(4):2405-2415. [doi: 10.1109/tii.2018.2873186]
3. Fagherazzi G. Deep digital phenotyping and digital twins for precision health: Time to dig deeper. J Med Internet Res 2020 Mar 03;22(3):e16770 [FREE Full text] [doi: 10.2196/16770] [Medline: 32130138]
4. Fuller A, Fan Z, Day C, Barlow C. Digital twin: Enabling technologies, challenges and open research. IEEE Access 2020;8:108952-108971. [doi: 10.1109/access.2020.2998358]
5. Corral-Acero J, Margara F, Marciniak M, Rodero C, Loncaric F, Feng Y, et al. The 'Digital Twin' to enable the vision of precision cardiology. Eur Heart J 2020 Dec 21;41(48):4556-4564 [FREE Full text] [doi: 10.1093/eurheartj/ehaa159] [Medline: 32128588]
6. Shen C, He J, Huang Y. Discriminative correlation filter network for robust landmark tracking in ultrasound guided intervention. 2019 Presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention; October 13-17, 2019; Shenzhen, China p. 646-654 URL: https://doi.org/10.1007/978-3-030-32254-0_72 [doi: 10.1007/978-3-030-32254-0_72]
7. Wen T, Wang C, Zhang Y, Zhou S. A novel ultrasound probe spatial calibration method using a combined phantom and stylus. Ultrasound Med Biol 2020 Aug;46(8):2079-2089 [FREE Full text] [doi: 10.1016/j.ultrasmedbio.2020.03.018] [Medline: 32446677]

8. Huang T, Li R, Li Y, Zhang X, Liao H. Augmented reality-based autostereoscopic surgical visualization system for telesurgery. Int J Comput Assist Radiol Surg 2021 Nov 07;16(11):1985-1997. [doi: 10.1007/s11548-021-02463-5] [Medline: 34363583]
9. Tholey G, Desai J, Castellanos A. Force feedback plays a significant role in minimally invasive surgery: results and analysis. Ann Surg 2005 Jan;241(1):102-109. [doi: 10.1097/01.sla.0000149301.60553.1e] [Medline: 15621997]
10. Dang LM, Piran M, Han D, Min K, Moon H. A survey on Internet of Things and cloud computing for healthcare. Electronics 2019 Jul 09;8(7):768 [FREE Full text] [doi: 10.3390/electronics8070768]
11. Dey N, Ashour AS, Shi F, Fong SJ, Tavares JMRS. Medical cyber-physical systems: A survey. J Med Syst 2018 Mar 10;42(4):74. [doi: 10.1007/s10916-018-0921-x] [Medline: 29525900]
12. Kocabas O, Soyata T, Aktas MK. Emerging security mechanisms for medical cyber physical systems. IEEE/ACM Trans. Comput. Biol. and Bioinf 2016 May 1;13(3):401-416. [doi: 10.1109/tcbb.2016.2520933]
13. Bruynseels K, Santoni de Sio F, van den Hoven J. Digital twins in health care: Ethical implications of an emerging engineering paradigm. Front Genet 2018;9:31 [FREE Full text] [doi: 10.3389/fgene.2018.00031] [Medline: 29487613]

Abbreviations

CAS: computer-assisted surgery
CT: computed tomography
DT: digital twin
MRI: magnetic resonance imaging

Edited by C Lovis; submitted 22.11.21; peer-reviewed by S Mallick; comments to author 16.10.22; revised version received 23.10.22; accepted 25.10.22; published 08.11.22.

Please cite as:
Qin J, Wu J
Realizing the Potential of Computer-Assisted Surgery by Embedding Digital Twin Technology
JMIR Med Inform 2022;10(11):e35138
URL: https://medinform.jmir.org/2022/11/e35138
doi:10.2196/35138
PMID:36346669

Original Paper

Perspective Toward Machine Learning Implementation in Pediatric Medicine: Mixed Methods Study

Natasha Alexander[1], MBBS, MSc; Catherine Aftandilian[2], MD; Lin Lawrence Guo[3], PhD; Erin Plenert[3], MSc; Jose Posada[4], PhD; Jason Fries[5], PhD; Scott Fleming[5], BSc; Alistair Johnson[3], PhD; Nigam Shah[5], MD, PhD; Lillian Sung[1,3], MD, PhD

[1]Division of Haematology/Oncology, The Hospital for Sick Children, Toronto, ON, Canada

[2]Division of Hematology/Oncology, Department of Pediatrics, Stanford University, Palo Alto, CA, United States

[3]Program in Child Health Evaluative Sciences, Peter Gilgan Centre for Research and Learning, Toronto, ON, Canada

[4]Department of Systems Engineering and Computing, Universidad del Norte, Barranquilla, Colombia

[5]Stanford Center for Biomedical Informatics Research, Department of Biomedical Data Science, Stanford University, Palo Alto, CA, United States

Corresponding Author:
Lillian Sung, MD, PhD
Division of Haematology/Oncology
The Hospital for Sick Children
555 University Avenue
Toronto, ON, M5G1X8
Canada
Phone: 1 416 813 5287
Fax: 1 416 813 5979
Email: lillian.sung@sickkids.ca

Abstract

Background: Given the costs of machine learning implementation, a systematic approach to prioritizing which models to implement into clinical practice may be valuable.

Objective: The primary objective was to determine the health care attributes respondents at 2 pediatric institutions rate as important when prioritizing machine learning model implementation. The secondary objective was to describe their perspectives on implementation using a qualitative approach.

Methods: In this mixed methods study, we distributed a survey to health system leaders, physicians, and data scientists at 2 pediatric institutions. We asked respondents to rank the following 5 attributes in terms of implementation usefulness: the clinical problem was common, the clinical problem caused substantial morbidity and mortality, risk stratification led to different actions that could reasonably improve patient outcomes, reducing physician workload, and saving money. Important attributes were those ranked as first or second most important. Individual qualitative interviews were conducted with a subsample of respondents.

Results: Among 613 eligible respondents, 275 (44.9%) responded. Qualitative interviews were conducted with 17 respondents. The most common important attributes were risk stratification leading to different actions (205/275, 74.5%) and clinical problem causing substantial morbidity or mortality (177/275, 64.4%). The attributes considered least important were reducing physician workload and saving money. Qualitative interviews consistently prioritized implementations that improved patient outcomes.

Conclusions: Respondents prioritized machine learning model implementation where risk stratification would lead to different actions and clinical problems that caused substantial morbidity and mortality. Implementations that improved patient outcomes were prioritized. These results can help provide a framework for machine learning model implementation.

(JMIR Med Inform 2022;10(11):e40039) doi:10.2196/40039

KEYWORDS

machine learning; clinical utilization; preferences; qualitative interviews

Introduction

Machine learning has had growing popularity in clinical settings related to the widespread adoption of electronic health records [1-3], combined with increasing data storage and computational ability [4]. In this setting, machine learning can be useful for multiple purposes including (1) to facilitate diagnoses, as in pathology [5,6] and radiology [7]; (2) to make predictions about outcomes for risk stratification; and (3) to improve resource utilization by anticipating volumes of patients or services [8]. However, despite the initial enthusiasm around machine learning in health care, domain experts have expressed caution [9,10]. Similar information technology solutions have commonly failed to be implemented or provide utility [11].

An important consideration impacting utility is choosing the clinical setting and problem in which a machine learning model is to be implemented [11]. A machine learning model's predictions need to augment current approaches in a way that is meaningful and actionable without introducing excessive burden. It is important to carefully plan a machine learning model's implementation because the costs of model deployment are considerable. Such costs may include resources required to develop and maintain the machine learning model, training of the intended model users regarding how to access and interpret the model's predictions, and support to help users implement the results into practice [12,13].

Given these costs, a systematic approach for determining which machine learning models should be prioritized for implementation into clinical practice may be valuable. In determining priorities, it would be important to involve key stakeholders at the institution in which deployment is planned. We chose to survey 2 pediatric centers, 1 in the United States with a more established biomedical informatics program, and 1 in Canada with a less established biomedical informatics program, to gain insight into whether experience and expertise affected preferences for machine learning model prioritization. Consequently, the primary objective was to determine the health care attributes respondents at 2 pediatric institutions rate as important when prioritizing machine learning model implementation. The secondary objective was to describe their perspectives on machine learning model implementation using a qualitative approach.

Methods

Study Design and Setting

This was a mixed methods study that included a quantitative and a qualitative component. The institutions were The Hospital for Sick Children (SickKids) in Toronto, Ontario, Canada, and Lucile Packard Children's Hospital in Palo Alto, California, United States.

Participants

We included health system leaders, physicians, and data scientists at SickKids and Lucile Packard Children's Hospital at the time of survey distribution. We excluded trainees.

Procedures

The survey was developed by the study team based on their impression of health care attributes respondents might consider to be important; the machine learning–focused questions are presented as Multimedia Appendix 1. Potential participants were identified through organizational emailing lists. The quantitative survey was distributed by email and participants completed the survey in REDCap [14]. The survey asked respondents to indicate whether they were health system leaders, physicians, or data scientists; respondents could indicate multiple categories. Demographic variables included clinical specialty (if applicable), years employed following completion of training, and gender.

We then asked about their knowledge of artificial intelligence on a 5-point Likert scale ranging from 1 (no knowledge at all) to 5 (a lot of knowledge). We asked them to rate their understanding of how machine learning models are built and interpreted, and how statistics are conducted and interpreted, using 5-point Likert scales ranging from 1 (no understanding) to 5 (fully understand). We asked if they had decision-making ability to implement artificial intelligence initiatives within their work environment, and how many machine learning models had been deployed at their institutions in the last 5 years.

The next section asked respondents to rank the following 5 clinical problem and implementation consequence attributes in terms of whether machine learning implementation would be useful: "the clinical problem being solved is common," "the clinical problem causes substantial morbidity or mortality," "risk stratification would lead to different clinical actions that could reasonably improve patient outcomes," "implementing the model could reduce physician workload," and "implementing the model could save money." Important attributes were defined as those ranked as most important or second most important (rank of 1 or 2) by respondents. The survey then asked 2 open-ended questions focused on clinical areas where being able to accurately predict an outcome might be useful, and clinical areas in which prioritization or reorganization of waitlists might be useful. Finally, the survey asked whether they would be willing to participate in a qualitative interview.

For the qualitative aspect, we purposively sampled respondents to maximize variation by institution and self-rated understanding of machine learning. Semistructured interviews were conducted using Zoom (Zoom Video Communications, Inc.) or Microsoft Teams by a member of the SickKids team (EP) with expertise in the conduct of qualitative interviews. Respondents were asked to list 3 scenarios in which a machine learning model for risk stratification could be useful and then to state which scenario was the most important to implement first and the rationale for the choice. They were then asked how they would feel about using a machine learning model for risk stratification as opposed to their current approach, and to describe concerns they had about using a machine learning model to guide patient care. The interviews were recorded and transcribed verbatim.

Analysis

The data from the quantitative survey from SickKids and Lucile Packard Children's Hospital were compared using the Fisher exact test. Analyses were performed in R (R Core Team) using RStudio version 3.6.1 [15,16].

The analysis of qualitative data was performed according to the principles of grounded theory methodology; data collection and analysis occurred concurrently. Qualitative transcripts were analyzed by 2 independent reviewers (NA and EP) using the constant comparative method to develop a theoretical framework for respondents' perspectives of machine learning that are grounded in their individual experiences and understandings. Sampling was continued until saturation was reached, which was defined as the point in which no new themes emerged from the data.

Ethics Approval

The study was approved by the Research Ethics Board at SickKids. The need for Institutional Review Board approval was waived by Lucile Packard Children's Hospital as the data collection was performed by SickKids personnel. For the quantitative survey, completion of the survey was considered implied consent to study participation. For the qualitative component, respondents provided verbal consent to participate.

Results

The quantitative survey was distributed at SickKids between November 1, 2021, and January 6, 2022 and at Lucile Packard Children's Hospital between March 15, 2022, and April 12, 2022. Among 613 eligible respondents, 275 (44.9%) responded. Figure 1 shows the participant identification and selection flowchart, including the number participating in the qualitative interviews when saturation was reached.

Table 1 presents the demographic characteristics of respondents; physician specialty (*P*<.001) and years from completion of training (*P*=.006) were significantly different between the 2 institutions. The majority of respondents were physicians (165/195, 84.6%, at SickKids and 73/80, 91.3%, at Lucile Packard Children's Hospital). The number of respondents who had decision-making ability to implement artificial intelligence initiatives was 99/195 (50.8%) at SickKids and 41/80 (51.3%) at Lucile Packard Children's Hospital. Most respondents did not know the number of machine learning models deployed at their institution over the last 5 years (137/195, 70.3%, at SickKids and 53/80, 66.3%, at Lucile Packard Children's Hospital).

Table 2 illustrates respondents' self-perceived knowledge of artificial intelligence and understanding of machine learning and statistics. There were no statistically significant differences in these ratings by institution (artificial intelligence knowledge, *P*=.93; machine learning development and interpretation, *P*=.72; statistics conduct and interpretation, *P*=.19). The percentage of respondents who stated they had "moderate" or "a lot" of artificial intelligence knowledge was 17.9% (35/195) at SickKids and 17.5% (14/80) at Lucile Packard Children's Hospital. Multimedia Appendix 2 compares respondent characteristics by those who self-rated their artificial intelligence knowledge as high (score of 4 or 5 on the 5-point Likert scale) versus not high across institutions. Those who self-rated their knowledge as high were significantly more likely to be males (*P*=.02) and nonphysicians (*P*=.006). The percentage of respondents who stated they understood machine learning development and interpretation at a "moderate" level or "fully" was 15.9% (31/195) at SickKids and 11.3% (9/80) at Lucile Packard Children's Hospital. Across both institutions, the percentage who stated their understanding of machine learning was "none" or "very little" was 146/275 (53.1%). Conversely, the percentage of respondents who stated they understood statistics conduct and interpretation at a "moderate" level or "fully" was 54.4% (106/195) at SickKids and 42.5% (34/80) at Lucile Packard Children's Hospital. Across both institutions, the percentage who stated their understanding of statistics was "none" or "very little" was 30/275 (10.9%).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram of participant identification, selection, and participation.

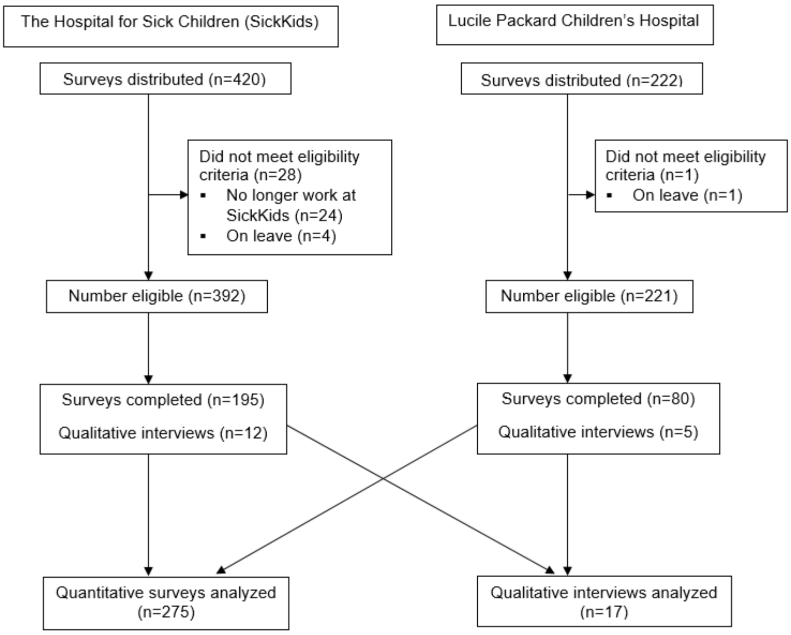

The Hospital for Sick Children (SickKids)
Lucile Packard Children's Hospital
Surveys distributed (n=420)
Surveys distributed (n=222)
Did not meet eligibility criteria (n=28)
▪ No longer work at SickKids (n=24)
▪ On leave (n=4)
Did not meet eligibility criteria (n=1)
▪ On leave (n=1)
Number eligible (n=392)
Number eligible (n=221)
Surveys completed (n=195)
Qualitative interviews (n=12)
Surveys completed (n=80)
Qualitative interviews (n=5)
Quantitative surveys analyzed (n=275)
Qualitative interviews analyzed (n=17)

Table 1. Demographic characteristics of participants at 2 pediatric institutions (N=275).

Characteristic	SickKids (n=195), n (%)	Lucile Packard Children’s Hospital (n=80), n (%)	*P* value
Male gender	93 (47.7)	35 (43.8)	.64
Professional role[a]			
Physician	165 (84.6)	73 (91.3)	.20
Health system leader	22 (11.3)	17 (21.3)	.05
Data scientist	15 (7.7)	2 (2.5)	.18
Physician specialty			<.001
Hematology oncology	33 (16.9)	14 (17.5)	
General medicine	21 (10.8)	7 (8.8)	
Critical care medicine	11 (5.6)	12 (15.0)	
Emergency medicine	14 (7.2)	0 (0)	
Cardiology	9 (4.6)	7 (8.8)	
Neurology	11 (5.6)	3 (3.8)	
Endocrinology and metabolism	10 (5.1)	6 (7.5)	
Gastroenterology	9 (4.6)	0 (0)	
Respirology	4 (2.1)	4 (5.0)	
Infectious disease	2 (1.0)	5 (6.3)	
Surgery	0 (0)	6 (7.5)	
Adolescent medicine	6 (3.1)	0 (0)	
Other	20 (10.3)	7 (8.8)	
Not known	45 (23.1)	9 (11.3)	
Years from completion of training			.006
<1	6 (3.1)	0 (0)	
1-4	38 (19.5)	5 (6.3)	
5-10	38 (19.5)	25 (31.3)	
11+	113 (57.9)	50 (62.5)	
Decision-making ability to implement artificial intelligence initiatives	99 (50.8)	41 (51.3)	>.99
Number of machine learning models deployed at institution in last 5 years			.43
None	31 (15.9)	11 (13.8)	
1	7 (3.6)	6 (7.5)	
2-4	14 (7.2)	9 (11.3)	
5-10	2 (1.0)	1 (1.3)	
11+	4 (2.1)	0 (0)	
Do not know	137 (70.3)	53 (66.3)	

[a] Respondent may choose more than 1 option and thus, numbers do not add to 100%.

Table 2. Self-rating of knowledge of artificial intelligence and understanding of machine learning and statistics.

Areas	SickKids (n=195), n (%)	Lucile Packard Children's Hospital (n=80), n (%)	*P*-value
Artificial intelligence knowledge			.93
None	10 (5.1)	5 (6.3)	
Very little	67 (34.4)	30 (37.5)	
Some	83 (42.6)	31 (38.8)	
Moderate	30 (15.4)	11 (13.8)	
A lot	5 (2.6)	3 (3.8)	
Machine learning development and interpretation			.72
None	44 (22.6)	18 (22.5)	
Very little	56 (28.7)	28 (35.0)	
Somewhat	64 (32.8)	25 (31.3)	
Moderate	24 (12.3)	8 (10.0)	
Fully	7 (3.6)	1 (1.3)	
Statistics conduct and interpretation			.19
None	4 (2.1)	1 (1.3)	
Very little	18 (9.2)	7 (8.8)	
Somewhat	67 (34.4)	38 (47.5)	
Moderate	78 (40.0)	29 (36.3)	
Fully	28 (14.4)	5 (6.3)	

Table 3 reveals the proportion of respondents who ranked each attribute as important (ranked first or second among the 5 attributes) for prioritization of machine learning models. There were no significant differences in these proportions by institution for any of the 5 attributes (Table 3). Across both sites, the most common important attributes were risk stratification leading to different actions (205/275, 74.5%) and clinical problem causes substantial morbidity or mortality (177/275, 64.4%). The attributes considered least important were "implementing the model could reduce physician workload" (40/275, 14.5%) and "implementing the model could save money" (13/275, 4.7%). The median importance scores for both institutions combined are also shown in Table 3 (where lower is more important).

Table 3. Ranked as important[a] by respondents for prioritization of machine learning.

Attributes considered important	SickKids (n=195), n (%)	Lucile Packard Children's Hospital (n=80), n (%)	*P*-value	Median importance score (IQR)[b]
The clinical problem being solved is common	66 (33.8)	35 (43.8)	.16	3 (2-3)
The clinical problem causes substantial morbidity or mortality	133 (68.2)	44 (55.0)	.05	2 (2-3)
Risk stratification would lead to different clinical actions that could reasonably improve patient outcomes	145 (74.4)	60 (75.0)	>.99	1 (1-2)
Implementing the model could reduce physician workload	29 (14.9)	11 (13.8)	.96	4 (3-4)
Implementing the model could save money	11 (5.6)	2 (2.5)	.42	5 (4-5)

[a] Important defined as attributes ranked as most important or second most important (rank of 1 or 2) in terms of whether a machine learning model would be useful.

[b] Across both institutions.

Table 4 shows the themes and subthemes from the qualitative interviews. Perceived benefits of machine learning model implementation included facilitating decision making in complex scenarios, supporting less experienced clinicians, reducing cognitive load, and reducing cognitive bias. It was also expressed that machine learning models can potentially improve the quality of care through standardization, more effective triage, and facilitating precision medicine. Finally, machine learning models had the potential to reduce physician workload. However, perceived challenges of machine learning model implementation included the potential for algorithmic bias, lack of transparency and trust, and failure to incorporate clinical expertise. Machine learning model implementation might also adversely affect quality of care and respondents spoke about the need to evaluate the impact of machine learning model implementation. Practical concerns raised about machine learning model implementation included challenges incorporating the model into the clinical workflow and questions about accountability in the event of poor outcomes arising from machine learning model–directed actions. Finally, uncertainty about the physician's role was identified. When asked to prioritize 1 clinical scenario for machine learning model implementation, the rationale for choosing which scenario to implement consistently related to impact on patient outcomes: "most benefit to kids," "leading cause of death," and "implications can be extremely serious."

Multimedia Appendix 3 illustrates examples of clinical areas that could be prioritized for machine learning initiatives identified from the quantitative survey.

Table 4. Perspectives of machine learning implementation in pediatric medicine from qualitative interviews.

Themes and subthemes	Example quotations
Benefits of machine learning implementation	
Facilitates decision making	
Complex scenario	*To me was very disturbing scenario where a very complex child with a number of issues, [...] Having some kind of system which alerts physicians who are directly involved as to not any in their own domains, but in other domains' risk would be helpful*
Support less experienced clinicians	*Well, you know where I see potential strength is not so much for the highly experienced physician, but more for the person who's starting out [...] and just doesn't have that experience base yet.*
Reduce cognitive load	*It can offload some of the cognitive load. So yeah, absolutely. I mean there's many times you find yourself in the middle of the night very tired, half groggy and trying to make a decision and kind of going back and forth in your brain. You know, for like half an hour - should I do this or that?*
Reduce cognitive bias	*[...] it's not that it replaces your judgment, it supplements another sense. ... your decisions informed no matter by your experience but it's informed by thousands of experiences, computed even more times to see all the possibilities and then come up with a best sort of path forward. The most likely scenario. And understanding that it is not a perfect prediction but it's a much more ... It's where that big data come in, right? It's really powered by real knowledge. It's not personal perceptions or personal experience, which is very biased and skewed.*
Improve quality of care	
Standardize care	*There probably is some significant interpersonal variability in terms of interpreting the guidelines and then decision making around management, and so if we could use machine learning so that there's less of that, all the while providing I guess more accurate or better care. I think that would be very helpful.*
More effective triage	*I feel like if we were able to use machine learning to risk stratify so that kids who are at higher risk could get more timely access to a referral. Recognizing that in this particular situation, certainly early diagnosis and management can really impact the trajectory of a child's outcome. I think that would be helpful.*
Facilitate precision medicine	*And what I mean by that if you look at it, look at a population of babies who were all born, say at 25 weeks. There will be individual differences that should [...] be detectable by machine learning or artificial intelligence. So instead of treating every baby as simply a member of the population, I can sort of drill down onto specific physiological and clinical factors for that baby, [...] get closer to the idea of personalized medicine.*
Reduce physician workload	
Freeing up time for physicians	*If it was really useful, then maybe it would free me up to do things that only I can do.*
Challenges with machine learning implementation	
Hinders decision making	
Algorithmic bias	*It's all about the biases like built into the system and how it's learned the data that you're putting in, and then how you get that out and how it would either pick up on our own biases, or like pre-existing, whether those are like systemic like sort of racial, ethnic or gendered biases [...] And so then that's not really helping us.*
Lack of transparency and trust	*Understanding what it is doing: like if it's doing things that I can't follow or don't understand, I'm going to be less to trust its opinion [...] I want to understand how it came to that decision so I can ask myself if I agree.*
Not incorporating clinical expertise into decisions	*I think it's like all the tools we have in medicine that if you use it appropriately, it can be incredibly powerful. But if it's used as a, you know, let me abandon all my other skills and I'll just follow this kind of direction, it potentially could be harmful, so I think a lot of thought will be needed.* *I mean in some ways it helps to predict, but I think I've always been a little skeptical about machine learning because biology and people do not follow an algorithm, they don't follow a formula.*
Negative impact on quality of care	
Need for outcome evaluation	*[...] looking at what the outcomes are and that we're actually improving patient care. So if we're admitting more but the outcomes are the same and the return visits are the same, then did it really matter and are we improving patient care or we just increasing cost to the system? And so, I think it needs constant evaluation, just like anything else that we do...*
Data quality	*Of course, you know your outcome or the recommendation, or how machine learning is used is always only as good as the input, right?*
Practical concerns	
Challenges in workflow implementation	*I guess there's going to be some learning curve. How do we use it? Is it feasible? Is it on my iPhone? Do I have to go into certain area, how fast will it take me to get the response and along with the interface, how friendly is the interface? You know things that are related to stuff that we have not seen yet.*

Themes and subthemes	Example quotations
Accountability	*The challenge with machine learning over clinical decision rules is right now with the accountability piece and it's just getting to what that's going to be like. We don't blame, you know, the lab test or the lab. You know, if we don't pick it up. But right now, I think people feeling if they go against it, what does that mean and do we have to add like admit everybody or treat everybody based on that, knowing that like you alluded on the first question that it is a probability [...] So what does that mean for the provider thing choose to ignore it versus if they choose to follow it in harm happens*
Physician role	
Uncertainty in physician role	*On the other hand, you know, maybe it also kind of takes away a little bit from like, I guess there's a fear of what exactly is the doctor's role. If the computer can do a better job at diagnosing then I can*

Discussion

In this mixed methods study, we found that the attributes most commonly listed as important for machine learning model implementation were risk stratification leading to different actions that could reasonably improve patient outcomes and a clinical problem that causes substantial morbidity or mortality. Few respondents considered reducing physician workload and saving money as important. We also found that important attributes were similar at the 2 institutions despite different levels of biomedical informatic program establishment and different health care systems.

The wide range of recommended areas for machine learning model implementation highlights the need for prioritization given the likely limited capacity to develop, deploy, and monitor machine learning models, even at large institutions with mature bioinformatics programs. This study is important as it provides a framework by which institutional leaders could make decisions about which machine learning models to prioritize for implementation. While we found that risk stratification that improves patient outcomes was the most common important attribute, additional considerations include actions that would arise from high- and low-risk labels, evidence that differential actions will improve outcomes, and identifying ideal thresholds for risk categorization. Even once a model is deployed, ongoing monitoring of model performance and the impact of model deployment on patient care and clinical workflows are additional postimplementation considerations.

While we evaluated attribute importance across respondent types, Wears and Berg [11] previously discussed the complex relationship between decision makers, beneficiaries of a machine learning solution, and those who shoulder the burden of implementation. They noted that a mismatch between these individuals can lead to failure. More specifically, it is often the administrator who is the decision maker and recipient of benefits, while it is the clinician who often shoulders the burden of implementation [11]. Anticipation and acknowledgement of conflicting perspectives will be required during the prioritization process among stakeholder types.

We also found that across both institutions, respondents had greater confidence in their understanding of statistics and relatively lower confidence in their understanding of machine learning. These perspectives did not differ between the 2 institutions despite different levels of establishment of their biomedical informatic programs. Our results suggest that across pediatric medicine in general, more education focused on machine learning is required during training and continuing education.

Our results complement the work of others who have highlighted the requirements of clinical decision support including those based on machine learning. Items important to consider include the need to avoid black boxes, excessive time requirement, and complexity in addition to ensuring relevance, respect, and scientific validity [17-19]. It also accompanies work demonstrating that barriers to adoption of artificial intelligence are not restricted to clinicians but also include parents [20,21]. It may also be useful to compare our findings with studies conducted outside of pediatric medicine. We found that the main anticipated benefits of machine learning implementation were facilitation of decision making, improvement in quality of care, and reduction in physician workload. Compared with our findings, benefits and challenges associated with artificial intelligence were similar in ophthalmology, dermatology, radiology, optometry, and surgery [22,23]. However, our study is unique because of the consideration of how to prioritize problems for implementation, a pragmatic consideration in developing a clinical program. In addition, the focus on pediatrics may be important as the nature of clinical problems, perspectives, and stakeholders can differ between pediatric and adult patient populations.

The strengths of this study include its mixed methods design and inclusion of 2 different pediatric institutions by country and establishment of their biomedical informatic programs. However, our results should be interpreted in light of their limitations. We had a relatively low response rate; respondents were likely biased in favor of interest in machine learning. Thus, nonrespondents likely would have had lower familiarity with machine learning and likely would have had less strong opinions about attributes considered important for machine learning prioritization. We also had a greater proportion of physicians than system leaders or data scientists; these groups may have different priorities or implementation concerns.

In conclusion, respondents prioritized machine learning model implementation where risk stratification would lead to different actions and clinical problems that caused substantial morbidity and mortality. Implementations that improved patient outcomes were prioritized. These results can help provide a framework for prioritizing machine learning model implementation.

Acknowledgments

LS is supported by the Canada Research Chair in Pediatric Oncology

Conflicts of Interest

None declared.

Multimedia Appendix 1
Quantitative survey administered.
[DOCX File , 16 KB - medinform_v10i11e40039_app1.docx]

Multimedia Appendix 2
Comparison of participants with artificial intelligence knowledge high versus not high (N=275).
[DOCX File , 17 KB - medinform_v10i11e40039_app2.docx]

Multimedia Appendix 3
Examples of recommendations of areas in pediatric care that should be prioritized for machine learning from quantitative survey.
[DOCX File , 18 KB - medinform_v10i11e40039_app3.docx]

References

1. Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. JMIR Med Inform 2014 Jan 17;2(1):e1 [FREE Full text] [doi: 10.2196/medinform.2913] [Medline: 25600256]
2. Hansen MM, Miron-Shatz T, Lau AYS, Paton C. Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. Yearb Med Inform 2018 Mar 05;23(01):21-26. [doi: 10.15265/iy-2014-0004]
3. El Aboudi N, Benhlima L. Big Data Management for Healthcare Systems: Architecture, Requirements, and Implementation. Adv Bioinformatics 2018 Jun 21;2018:4059018-4059010 [FREE Full text] [doi: 10.1155/2018/4059018] [Medline: 30034468]
4. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019 Apr 04;380(14):1347-1358. [doi: 10.1056/nejmra1814259]
5. Nanaa A, Akkus Z, Lee WY, Pantanowitz L, Salama ME. Machine learning and augmented human intelligence use in histomorphology for haematolymphoid disorders. Pathology 2021 Apr;53(3):400-407. [doi: 10.1016/j.pathol.2020.12.004] [Medline: 33642096]
6. Stenzinger A, Alber M, Allgäuer M, Jurmeister P, Bockmayr M, Budczies J, et al. Artificial intelligence and pathology: From principles to practice and future applications in histomorphology and molecular profiling. Semin Cancer Biol 2022 Sep;84:129-143. [doi: 10.1016/j.semcancer.2021.02.011] [Medline: 33631297]
7. Richardson ML, Adams SJ, Agarwal A, Auffermann WF, Bhattacharya AK, Consul N, et al. Review of Artificial Intelligence Training Tools and Courses for Radiologists. Acad Radiol 2021 Sep;28(9):1238-1252. [doi: 10.1016/j.acra.2020.12.026] [Medline: 33714667]
8. Drysdale E, Singh D, Goldenberg A. Forecasting emergency department capacity constraints for COVID isolation beds. arXiv. Preprint posted online on November 9, 2020 2020;1:1 [FREE Full text]
9. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. JAMA 2019 Jan 01;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
10. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA 2018 Jan 02;319(1):19-20. [doi: 10.1001/jama.2017.19198] [Medline: 29261830]
11. Wears RL, Berg M. Computer technology and clinical work: still waiting for Godot. JAMA 2005 Mar 09;293(10):1261-1263. [doi: 10.1001/jama.293.10.1261] [Medline: 15755949]
12. Morse KE, Bagley SC, Shah NH. Estimate the hidden deployment cost of predictive models to improve patient care. Nat Med 2020 Jan 13;26(1):18-19. [doi: 10.1038/s41591-019-0651-8] [Medline: 31932778]
13. Sendak M, Balu S, Schulman K. Barriers to Achieving Economies of Scale in Analysis of EHR Data. Appl Clin Inform 2017 Dec 20;08(03):826-831. [doi: 10.4338/aci-2017-03-cr-0046]
14. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
15. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. URL: https://www.R-project.org/
16. RStudio Team. RStudio: Integrated Development for R. Boston, MA: RStudio, PBC; 2020. URL: http://www.rstudio.com/
17. Rousseau N, McColl E, Newton J, Grimshaw J, Eccles M. Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. BMJ 2003 Feb 08;326(7384):314 [FREE Full text] [doi: 10.1136/bmj.326.7384.314] [Medline: 12574046]

18. Zheng K, Padman R, Johnson MP, Diamond HS. Understanding technology adoption in clinical care: clinician adoption behavior of a point-of-care reminder system. Int J Med Inform 2005 Aug;74(7-8):535-543. [doi: 10.1016/j.ijmedinf.2005.03.007] [Medline: 16043083]
19. Patterson ES, Doebbeling BN, Fung CH, Militello L, Anders S, Asch SM. Identifying barriers to the effective use of clinical reminders: bootstrapping multiple methods. J Biomed Inform 2005 Jun;38(3):189-199 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.015] [Medline: 15896692]
20. Sisk BA, Antes AL, Burrous S, DuBois JM. Parental Attitudes toward Artificial Intelligence-Driven Precision Medicine Technologies in Pediatric Healthcare. Children (Basel) 2020 Sep 20;7(9):145 [FREE Full text] [doi: 10.3390/children7090145] [Medline: 32962204]
21. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, Utikal JS, Hauschild A, Schadendorf D, et al. Artificial Intelligence in Skin Cancer Diagnostics: The Patients' Perspective. Front Med (Lausanne) 2020 Jun 2;7:233 [FREE Full text] [doi: 10.3389/fmed.2020.00233] [Medline: 32671078]
22. Scheetz J, Rothschild P, McGuinness M, Hadoux X, Soyer HP, Janda M, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. Sci Rep 2021 Mar 04;11(1):5193 [FREE Full text] [doi: 10.1038/s41598-021-84698-5] [Medline: 33664367]
23. Eschert T, Schwendicke F, Krois J, Bohner L, Vinayahalingam S, Hanisch M. A Survey on the Use of Artificial Intelligence by Clinicians in Dentistry and Oral and Maxillofacial Surgery. Medicina (Kaunas) 2022 Aug 05;58(8):1059 [FREE Full text] [doi: 10.3390/medicina58081059] [Medline: 36013526]

Edited by C Lovis; submitted 02.06.22; peer-reviewed by S Ramgopal, H Hochheiser; comments to author 07.09.22; revised version received 15.09.22; accepted 10.10.22; published 17.11.22.

Please cite as:
Alexander N, Aftandilian C, Guo LL, Plenert E, Posada J, Fries J, Fleming S, Johnson A, Shah N, Sung L
Perspective Toward Machine Learning Implementation in Pediatric Medicine: Mixed Methods Study
JMIR Med Inform 2022;10(11):e40039
URL: https://medinform.jmir.org/2022/11/e40039
doi:10.2196/40039
PMID:36394938

Original Paper

The Use of Electronic Health Record Metadata to Identify Nurse-Patient Assignments in the Intensive Care Unit: Algorithm Development and Validation

Kathryn A Riman[1], PhD, RN; Billie S Davis[1], PhD; Jennifer B Seaman[2], PhD, RN, CHPN; Jeremy M Kahn[1], MD, MS

[1]Department of Critical Care Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States
[2]Department of Acute & Tertiary Care, University of Pittsburgh School of Nursing, Pittsburgh, PA, United States

Corresponding Author:
Jeremy M Kahn, MD, MS
Department of Critical Care Medicine
University of Pittsburgh School of Medicine
3550 Terrace Street
Pittsburgh, PA, 15261
United States
Phone: 1 412 683 7601
Email: jeremykahn@pitt.edu

Abstract

Background: Nursing care is a critical determinant of patient outcomes in the intensive care unit (ICU). Most studies of nursing care have focused on nursing characteristics aggregated across the ICU (eg, unit-wide nurse-to-patient ratios, education, and working environment). In contrast, relatively little work has focused on the influence of individual nurses and their characteristics on patient outcomes. Such research could provide granular information needed to create evidence-based nurse assignments, where a nurse's unique skills are matched to each patient's needs. To date, research in this area is hindered by an inability to link individual nurses to specific patients retrospectively and at scale.

Objective: This study aimed to determine the feasibility of using nurse metadata from the electronic health record (EHR) to retrospectively determine nurse-patient assignments in the ICU.

Methods: We used EHR data from 38 ICUs in 18 hospitals from 2018 to 2020. We abstracted data on the time and frequency of nurse charting of clinical assessments and medication administration; we then used those data to iteratively develop a deterministic algorithm to identify a single ICU nurse for each patient shift. We examined the accuracy and precision of the algorithm by performing manual chart review on a randomly selected subset of patient shifts.

Results: The analytic data set contained 5,479,034 unique nurse-patient charting times; 748,771 patient shifts; 87,466 hospitalizations; 70,002 patients; and 8,134 individual nurses. The final algorithm identified a single nurse for 97.3% (728,533/748,771) of patient shifts. In the remaining 2.7% (20,238/748,771) of patient shifts, the algorithm either identified multiple nurses (4,755/748,771, 0.6%), no nurse (14,689/748,771, 2%), or the same nurse as the prior shift (794/748,771, 0.1%). In 200 patient shifts selected for chart review, the algorithm had a 93% accuracy (ie, correctly identifying the primary nurse or correctly identifying that there was no primary nurse) and a 94.4% precision (ie, correctly identifying the primary nurse when a primary nurse was identified). Misclassification was most frequently due to patient transitions in care location, such as ICU transfers, discharges, and admissions.

Conclusions: Metadata from the EHR can accurately identify individual nurse-patient assignments in the ICU. This information enables novel studies of ICU nurse staffing at the individual nurse-patient level, which may provide further insights into how nurse staffing can be leveraged to improve patient outcomes.

(JMIR Med Inform 2022;10(11):e37923) doi:10.2196/37923

KEYWORDS

electronic health records; algorithms; intensive care unit; documentation; hospitals; workforce; nursing; clinical informatics; health informatics

Introduction

Critical care nurses encompass the single largest workforce in the intensive care unit (ICU) and provide essential patient care 24 hours a day, 7 days a week. Adequate nurse staffing is essential for high-quality critical care; a large body of literature shows an association between patient outcomes and nurse staffing patterns, including nurse-to-patient ratios, nurse education, and nurse work environments [1-7]. This literature has been instrumental in the development of ICU staffing guidelines that strengthen ICU nursing, leading to lower mortality in US hospitals [8,9]. As beneficial as these guidelines have been, one limitation is that they focus on ICU nurses on average, rather than as individuals with varying levels of expertise, experience, and familiarity with the other members of the interprofessional care team. As a result, these approaches fail to consider the specific actions and knowledge of individual critical care nurses at the bedside and fail to account for staffing changes that occur throughout a patient's ICU stay. These approaches are also subject to the ecological fallacy, since epidemiological relationships observed at the group level may not exist at the individual patient level [10].

More research is needed to understand how individual nurse characteristics, not just nursing characteristics in aggregate, influence patient outcomes. A critical barrier to progress in this area is the lack of a valid and reliable approach to link specific nurses to specific patients on a large scale. The electronic health record (EHR) is a potentially valuable resource for addressing this gap. Nurses use EHRs for a wide variety of tasks, including assessment documentation and medication administration. When completing these tasks, the nurse leaves behind metadata in the form of an electronic signature indicating that they were the person that performed the assessment or administered the medication. In theory, these metadata could be used to link individual nurses to specific patients during a shift, thereby generating a high-granularity measure of individual nurse-to-patient assignments. This approach would facilitate individual-level research examining the association between various nurse characteristics and patient outcomes. This research could also aid in the development of sophisticated algorithms that generate personalized nurse-to-patient assignments based on nurse skill and patient need. The objective of this study was to determine the feasibility of using the metadata from the EHR in the form of electronic signatures to determine nurse-patient assignments in the ICU.

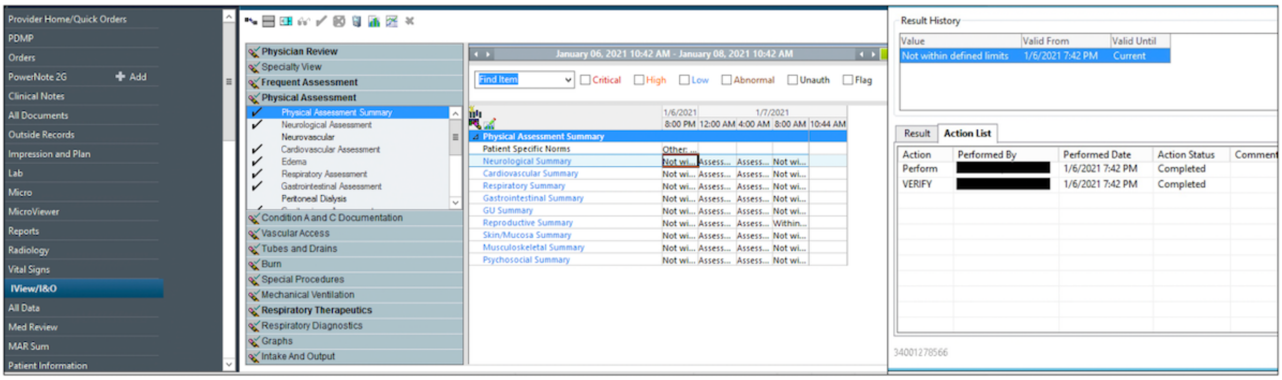
Methods

Study Design and Data

We developed and validated an algorithm for retrospectively linking individual nurses to individual patients at the level of the nursing shift. The study was conducted in a multihospital health care system in Western Pennsylvania in the United States. All hospitals shared a single enterprise-wide electronic medical record (Cerner PowerChart, Cerner Corporation) with all data warehoused in a single integrated database. All patient-level data and nurse metadata were obtained from this warehouse. To collect the data, key data elements were first identified by investigators with knowledge of the relevant clinical workflows. Relevant data were then extracted from the Cerner Millennium database (Oracle Cerner) using Cerner command language by a centralized research information technology team and transferred to the investigative team as text files (.txt) via Globus secure transfer. Data integrity was assessed for issues such as delimiter and string errors using Python (version 3.10.7; Python Software Foundation). The resulting text files were uploaded into a Microsoft SQL Server database (Microsoft Corp). Metadata of interest included date- and time-stamped electronic signatures on clinical assessments (eg, level of sedation, cardiac rhythm assessments, and neurological assessments) and medication administrations (Figure 1). Patient data included demographics, discharge disposition, as well as date and time stamps for admissions and discharges at the hospitalization and ICU-stay level. Patient data and nursing metadata were linked using direct patient identifiers.

Patients qualifying for inclusion in the analytic sample included adult patients admitted to 38 ICUs in 18 hospitals with discharge dates from January 1, 2018, to September 30, 2020. There were no specific exclusion criteria. We divided all ICU admissions between January 1, 2018, and August 31, 2020, into mutually exclusive 12-hour nursing shifts. We defined the day shift as 7 AM to 6:59:59 PM, and the night shift as 7 PM to 6:59:59 AM the following morning.

Figure 1. Sample screenshot of the location of nurse metadata in Cerner PowerChart. Higher-resolution version of this figure is available in Multimedia Appendix 1.



	1/6/2021	1/7/2021			
	8:00 PM	12:00 AM	4:00 AM	8:00 AM	10:44 AM
Physical Assessment Summary					
Patient Specific Norms	Other: ...				
Neurological Summary	Not wi...	Assess...	Assess...	Not wi...	
Cardiovascular Summary	Not wi...	Assess...	Assess...	Not wi...	
Respiratory Summary	Not wi...	Assess...	Assess...	Not wi...	
Gastrointestinal Summary	Not wi...	Assess...	Assess...	Not wi...	
GU Summary	Not wi...	Assess...	Assess...	Not wi...	
Reproductive Summary	Not wi...	Assess...	Assess...	Within...	
Skin/Mucosa Summary	Not wi...	Assess...	Assess...	Not wi...	
Musculoskeletal Summary	Not wi...	Assess...	Assess...	Not wi...	
Psychosocial Summary	Not wi...	Assess...	Assess...	Not wi...	

Value	Valid From	Valid Until
Not within defined limits	1/6/2021 7:42 PM	Current

Action	Performed By	Performed Date	Action Status	Comment
Perform	[redacted]	1/6/2021 7:42 PM	Completed	
VERIFY	[redacted]	1/6/2021 7:42 PM	Completed	

Algorithm Development and Validation

Our algorithm has 2 main input variables from the nurse metadata: (1) the count of the number of unique times a nurse charted per patient shift and (2) the length of time between the nurse's first and last charting times per patient shift. The count of the number of times a nurse charted per patient shift was used based on the assumption that a patient's primary nurse would chart more frequently than other nurses. We only counted unique times instead of all charting instances because nurses could electronically sign multiple charting instances at one time. The second input variable—the length of time between the nurse's first and last charting times per patient shift—was used based on the assumption that the primary nurse would have a longer interval between first and last charting times compared to other nurses.

Using these 2 input variables, we developed a 2-step algorithm with the following processing methods. In step 1 of the algorithm, for each patient shift, we identified the primary nurse as the nurse with the highest number of charting times during the shift, breaking ties by the longest interval between first and last charting times. If a tie remained (ie, there was more than one nurse with the same number of charting times and same charting interval), we considered there to be no primary nurse during that shift. We made this decision because we felt we could not further downselect without introducing randomness into the algorithm. In step 2, we repeated the method in step 1 but excluded the nurse from the current shift if they were the primary nurse in the prior shift (from step 1), based on the assumption that the algorithm might erroneously identify a nurse that performed an extensive amount of charting after their shift was complete. At the end of this process there were 2 output variables, as follows: (1) a binary variable indicating if each patient shift had either a primary nurse identified or not; and (2) the identity of the nurse for shifts in which a nurse was identified. Figure 2 depicts the logic model of the nurse-to-patient assignment algorithm.

To examine the underlying mechanism of the algorithm, we examined how each shift was either assigned a primary nurse or not assigned a primary nurse. Primary nurse assignment could occur in one of the following 3 ways: (1) only one nurse charted on the patient in the shift and thus had the highest number of charting times; (2) multiple nurses charted on the patient in the shift, but only one of them had the highest number of charting times; or (3) multiple nurses charted on the patient in the shift, more than one of them had the highest number of charting times, and the tie was broken based on the charting time interval. A shift could not be assigned a primary nurse also in one of the following 3 ways: (1) only one nurse charted on the patient in the shift, but it was the primary nurse in the prior shift; (2) multiple nurses charted on the patient in the shift, more than one of them had the highest number of charting times, and the tie was not broken based on the charting time interval; or (3) no nurses charted on the patient in the shift. Each shift was categorized into one of the above groups (Table 1).

We validated the performance of the nurse assignment algorithm against the reference standard of chart review. We selected a stratified random sample of 200 patient shifts, matching the proportion of patient shifts within each of the 6 categories described above. A nurse on the research team (KR) reviewed the charts, using the full range of clinical documentation to identify the actual primary nurse when such a nurse existed. We report the algorithm's performance based on its accuracy, defined as the sum of true positives and true negatives divided by the sum of true positives, true negatives, false positives, and false negatives; and precision, defined as true positives divided by the sum of true positives and false positives. We then performed an additional review of 50 randomly selected patient shifts where no primary nurse was identified. We used this review to supplement our understanding of the reasons why no primary nurse was identified.

We described the data set and the patient sample using standard summary statistics. Precision and accuracy are reported as proportions with exact 95% CIs calculated using the binomial distribution. Data management and statistical analyses were performed using Microsoft SQL Server and Stata (version 17.0; StataCorp).

Figure 2. Logic model of the nurse-to-patient assignment algorithm.

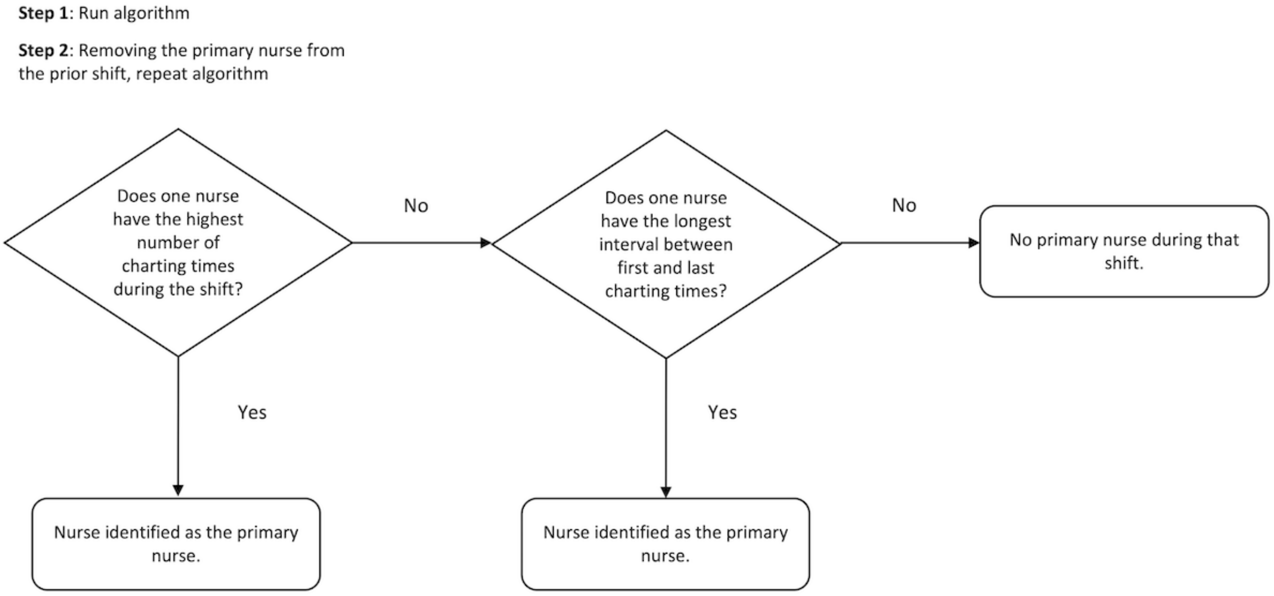

Step 1: Run algorithm
Step 2: Removing the primary nurse from the prior shift, repeat algorithm
Does one nurse have the highest number of charting times during the shift?
No
Does one nurse have the longest interval between first and last charting times?
No
No primary nurse during that shift.
Yes
Yes
Nurse identified as the primary nurse.
Nurse identified as the primary nurse.

Table 1. Algorithm results (N=748,771).

Characteristics	Patient shifts, n (%)
Primary nurse identified (n=728,533, 97.3%)	
One nurse charted	591,578 (79)
Multiple nurses charted but one nurse charted the most times	130,591 (17.4)
Multiple nurses charted the most times, tie broken by charting time interval	6364 (0.8)
Primary nurse not identified (n=20,238, 2.7%)	
No nurse charted	14,689 (2)
Multiple nurses charted the most times, tie not broken by charting time interval	4755 (0.6)
One nurse charted and it was the primary nurse in the prior shift	794 (0.1)

Ethics Approval

The University of Pittsburgh institutional review board approved the research protocol (19040420).

Results

The final analytic data file contained 5,479,034 nurse-patient charting times; 748,771 patient-shifts; 87,466 hospitalizations; and 70,002 patients (Table 2). There were 8,134 individual nurses in the data, with 4,797 (59.0%) of them identified as the primary nurse for at least one shift. Patients had a mean age of 63.8 (SD 17.1) years; 32,199 (46.%) were female; and 58,476 (83.5%) were White. Most patients were discharged to a long-term acute care hospital or skilled nursing facility (n=36,435, 52%) or home (n=22,380, 32%; Table 3).

The algorithm performance compared to the reference standard of chart review is reported in Table 4. The algorithm was highly accurate, correctly identifying the primary nurse or correctly identifying that there was no primary nurse 93% of the time. The algorithm was also quite precise, with 94.4% of cases having the correct primary nurse when a primary nurse was identified. In the few cases where the algorithm identified one primary nurse, but chart review identified a different primary nurse, it was typically due to either an operating room or floor nurse being identified, irregular shift lengths (eg, part time nurses), or emergent scenarios (eg, cardiac arrests) in which nurses shared tasks.

In the 5 cases from the main chart review where the algorithm did not identify a primary nurse and in the 50 supplemental chart review cases, we found that the underlying cause was due to a variety of circumstances. In about half of the cases, we could identify a primary nurse in chart review. However, the information was usually in elements of the EHR not visible to the algorithm, such as transfer or discharge forms, pain assessments, or arrangements after patient death. In other cases, chart review revealed that there were 2 primary nurses, as one of the nurses was being oriented to the unit. Finally, there were some cases where no nurse was identified even via chart review because there was no digital documentation to verify the identity of the primary nurse. This often occurred when the patient was admitted to the ICU late in the shift or discharged from the ICU early in the shift, such that the time spent in the ICU was very short.

Table 2. Data set characteristics.

Characteristics	Values, n
Number of hospitals	18
Number of intensive care units	38
Nurse-patient charting times	5,479,034
Patient shifts	748,771
Hospitalizations	87,466
Patients	70,002
Nurses	8134
Nurses ever identified as the primary nurse	4797

Table 3. Patient characteristics (N=70,002).

Characteristics	Values
Hospitalizations per patient, mean (SD); (min, max)	1.2 (0.8); (1, 33)
Shifts per patient, mean (SD); (min, max)	10.7 (16.4); (1, 561)
Age (years), mean (SD); (min, max)	63.8 (17.1); (18, 119)
Sex (female), n (%)	32,199 (46.0)
Race or ethnicity, n (%)	
White	58,476 (83.5)
Black	6816 (9.7)
Other	680 (1)
Missing	4030 (5.8)
Discharge disposition, n (%)	
Home	22,380 (32)
Transfer to short-term hospital	2087 (3)
Other transfer (LTAC[a], SNF[b])	36,435 (52)
Died	8024 (11.5)
Hospice	895 (1.3)
Other or missing	181 (0.3)

[a] LTAC: long-term acute care hospital.

[b] SNF: skilled nursing facility.

Table 4. Algorithm performance.

Characteristics	Same or newly identified primary nurse in chart review, n	Different or no primary nurse in chart review, n
Primary nurse from algorithm	184 (true positive)	11 (false positive)
No primary nurse from algorithm	3 (false negative)	2 (true negative)

Accuracy and precision were calculated as follows:

Looking back at the full data set, in the 97.3% (728,533/748,771) of patient shifts with a primary nurse identified, the median time in the ICU during the patient shift was 12 hours, compared to a median time in the ICU of 1.3 hours among the 2.7% (20,238/748,771) of patient shifts with a primary nurse not identified. In specific applications, researchers could exclude these shifts and expect an even stronger algorithm performance.

Discussion

We developed and validated an algorithm that identifies nurse-patient assignments using metadata from the EHR. Building on a body of literature linking hospital-level measures of nurse staffing to patient outcomes [11], this study presents a novel method for characterizing individual nurse-patient assignments. This method opens several new avenues of research into the influence of nurse staffing patterns on patient outcomes. With a direct linkage of nurse to patient, it will be possible to investigate the mechanisms, nursing characteristics, and team dynamics underlying the relationship between nurse staffing and patient outcomes. Our methodology can also be applied to other roles in the care team to investigate if similar associations are present.

More broadly, this study demonstrates the potential value of EHR metadata as a tool for understanding and improving health care delivery in the ICU. Existing publicly available data sets, such as Medical Information Mart for Intensive Care, use patient data from the EHR but do not contain information at the individual provider level or link those providers to patients [12]. Registered nurses provide bedside surveillance 24 hours a day, 7 days a week and are often the first members of the care team to recognize patient deterioration. By linking individual nurses to patients, our methods progress beyond unit-wide measures of staffing and nurse characteristics and allow the generation of more granular measures to study the relationship between nurse staffing and patient outcomes.

Our work builds off prior efforts that use EHR metadata to assess health care team structure and function [13-19]. Unlike those studies, our study focused on a specific provider type and used patient care–focused metadata rather than data less tightly linked to actual care, such as the data left when an electronic chart is accessed. Informed by prior work, our method could be applied to other roles within the health care team (eg, respiratory therapists and physical therapists) to examine and optimize team dynamics and collaboration [13,15,16]. Similar to work conducted by Hribar and colleagues [14] in outpatient clinics, we may be able to examine the timing and density of tasks in the EHR to optimize scheduling of various interventions (eg, spontaneous breathing trials).

The main strengths of this study include the use of a large, multicenter data set with varying ICU types, and the innovation inherent in developing a novel yet generalizable algorithm that links nurses to patients using the EHR. Along with these strengths, this study also had several limitations that may be sources of bias or imprecision. First, the metadata we obtained were limited to EHR documentation of clinical assessments and medication administration. We focused on these domains because we considered them to be most tightly linked to clinical care, and therefore, most indicative of the actual bedside nurse. However, nurses chart other information in the EHR, and it is conceivable that using additional sources of metadata could lead to a misidentification of the bedside nurse, thereby worsening algorithm performance. Since the vast majority of shifts included relevant metadata, we suspect that any bias was minimal and overall would serve to increase the accuracy and precision of the algorithm. We also used only EHR metadata and not data from other sources, such as bed-tracking data that might directly identify the bedside nurse. Although these data may more readily allow for an accurate and precise identification of the bedside nurse, we made this decision to make our algorithm maximally generalizable, since many hospitals do not use bed-tracking software, while an increasing number of hospitals use EHRs [20]. We also chose to retain all patient shifts, not limiting to those with 12 hours in the ICU. With less time in the ICU, there is less of a chance for the primary nurse to leave behind their digital signature and a higher likelihood of misidentification (eg, assigning the ward nurse). Excluding such shifts would likely improve our algorithm performance, but we felt keeping them makes our algorithm more generalizable. Finally, the algorithm was developed using EHR data from several ICUs belonging to a large hospital system in Western Pennsylvania, which may lack generalizability to other settings and hospital systems. However, these hospitals are diverse in terms of size and academic status, making them largely representative of the US health care system.

In future work, it may be possible to apply this algorithm to other roles within the care team, such as respiratory therapists and physical therapists. Ultimately, identifying links between individual providers and individual patients will open new lines of inquiry into how provider characteristics and team characteristics are associated with individual patient outcomes. Beyond creating evidence-based nurse-to-patient assignments where the nurse’s skills are matched to the patient’s needs, we can also intentionally construct the care team to maximize care continuity and team connectedness [21,22].

In conclusion, this algorithm can accurately identify nurse-patient assignments based on nurse documentation in the EHR. This algorithm can be used by researchers to generate data on nurse-patient assignments and answer questions related to nurse health services research at the patient level and nurse assignment level.

Acknowledgments

Funding for this study was provided by the National Institutes of Health (NIH) and principal investigator JMK (T32HL007820; R35HL144804). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

Each of the authors has made a substantial contribution to this study’s conception or design; acquisition, analysis, and interpretation of data; drafting the manuscript or revising it critically; and approval of the version submitted for review.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Higher resolution version of Figure 1. Sample screenshot of the location of nurse metadata in Cerner PowerChart.

[PNG File , 852 KB - medinform_v10i11e37923_app1.png]

References

1. Cho S, Hwang J, Kim J. Nurse staffing and patient mortality in intensive care units. Nurs Res 2008;57(5):322-330. [doi: 10.1097/01.NNR.0000313498.17777.71] [Medline: 18794716]
2. Kelly DM, Kutney-Lee A, McHugh MD, Sloane DM, Aiken LH. Impact of critical care nursing on 30-day mortality of mechanically ventilated older adults. Crit Care Med 2014;42(5):1089-1095. [doi: 10.1097/ccm.0000000000000127]
3. Manojlovich M, DeCicco B. Healthy work environments, nurse-physician communication, and patients' outcomes. Am J Crit Care 2007;16:54. [doi: 10.4037/ajcc2007.16.6.536]
4. Neuraz A, Guérin C, Payet C, Polazzi S, Aubrun F, Dailler F, et al. Patient mortality is associated with staff resources and workload in the ICU. Crit Care Med 2015;43(8):1587-1594. [doi: 10.1097/ccm.0000000000001015]
5. Stone P, Mooney-Kane C, Larson E, Horan T, Glance LG, Zwanziger J, et al. Nurse working conditions and patient safety outcomes. Med Care 2007 Jun;45(6):571-578. [doi: 10.1097/MLR.0b013e3180383667] [Medline: 17515785]
6. Tarnow-Mordi W, Hau C, Warden A, Shearer A. Hospital mortality in relation to staff workload: a 4-year study in an adult intensive-care unit. The Lancet 2000 Jul;356(9225):185-189. [doi: 10.1016/s0140-6736(00)02478-8]
7. West E, Barron DN, Harrison D, Rafferty AM, Rowan K, Sanderson C. Nurse staffing, medical staffing and mortality in intensive care: an observational study. Int J Nurs Stud 2014 May;51(5):781-794 [FREE Full text] [doi: 10.1016/j.ijnurstu.2014.02.007] [Medline: 24636667]
8. Weled BJ, Adzhigirey LA, Hodgman TM, Brilli RJ, Spevetz A, Kline AM, et al. Critical care delivery. Crit Care Med 2015;43(7):1520-1525. [doi: 10.1097/ccm.0000000000000978]
9. AACN standards for establishing and sustaining healthy work environments: a journey to excellence. American Association of Critical-Care Nurses. URL: https://www.aacn.org/~/media/aacn-website/nursing-excellence/healthy-work-environment/execsum.pdf [accessed 2022-10-27]
10. Piantadosi S, Byar D, Green S. The ecological fallacy. Am J Epidemiol 1988 May;127(5):893-904. [doi: 10.1093/oxfordjournals.aje.a114892] [Medline: 3282433]
11. Rae P, Pearce S, Greaves P, Dall'Ora C, Griffiths P, Endacott R. Outcomes sensitive to critical care nurse staffing levels: a systematic review. Intensive Crit Care Nurs 2021 Dec;67:103110. [doi: 10.1016/j.iccn.2021.103110] [Medline: 34247936]
12. Sauer C, Dam T, Celi L, Faltys M, de la Hoz MAA, Adhikari L, et al. Systematic review and comparison of publicly available ICU data sets-a decision guide for clinicians and data scientists. Crit Care Med 2022 Jun 01;50(6):e581-e588. [doi: 10.1097/CCM.0000000000005517] [Medline: 35234175]
13. Gray JE, Feldman H, Reti S, Markson L, Lu X, Davis RB, et al. Using digital crumbs from an electronic health record to identify, study and improve health care teams. AMIA Annu Symp Proc 2011;2011:491-500 [FREE Full text] [Medline: 22195103]
14. Hribar MR, Read-Brown S, Reznick L, Lombardi L, Parikh M, Yackel TR, et al. Secondary use of EHR timestamp data: validation and application for workflow optimization. AMIA Annu Symp Proc 2015;2015:1909-1917 [FREE Full text] [Medline: 26958290]
15. Durojaiye A, Levin S, Toerper M, Kharrazi H, Lehmann HP, Gurses AP. Evaluation of multidisciplinary collaboration in pediatric trauma care using EHR data. J Am Med Inform Assoc 2019 Jun 01;26(6):506-515 [FREE Full text] [doi: 10.1093/jamia/ocy184] [Medline: 30889243]
16. Chen Y, Lehmann CU, Hatch LD, Schremp E, Malin BA, France DJ. Modeling care team structures in the neonatal intensive care unit through network analysis of EHR audit logs. Methods Inf Med 2019 Nov 13;58(4-05):109-123 [FREE Full text] [doi: 10.1055/s-0040-1702237] [Medline: 32170716]
17. Kricke G, Carson M, Lee Y. Leveraging electronic health record documentation for failure mode and effects analysis team identification. J Am Med Inform Assoc 2017;24:294. [doi: 10.1093/jamia/ocw083]
18. Vawdrey D, Wilcox L, Collins S, Feiner S, Mamykina O, Stein D, et al. Awareness of the care team in electronic health records. Appl Clin Inform 2017 Dec 16;02(04):395-405. [doi: 10.4338/aci-2011-05-ra-0034]
19. Soulakis N, Carson M, Lee Y, Schneider DH, Skeehan CT, Scholtens DM. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. J Am Med Inform Assoc 2015 Mar;22(2):299-311 [FREE Full text] [doi: 10.1093/jamia/ocu017] [Medline: 25710558]
20. Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of electronic health records in U.S. hospitals. N Engl J Med 2009 Apr 16;360(16):1628-1638. [doi: 10.1056/nejmsa0900592]
21. Kelly Costa D, Liu H, Boltey EM, Yakusheva O. The structure of critical care nursing teams and patient outcomes: a network analysis. Am J Respir Crit Care Med 2020 Feb 15;201(4):483-485. [doi: 10.1164/rccm.201903-0543le]

22. Costa DK, Valley TS, Miller MA, Manojlovich M, Watson SR, McLellan P, et al. ICU team composition and its association with ABCDE implementation in a quality collaborative. J Crit Care 2018 Apr;44:1-6 [FREE Full text] [doi: 10.1016/j.jcrc.2017.09.180] [Medline: 28978488]

Abbreviations

EHR: electronic health record
ICU: intensive care unit

Edited by C Lovis; submitted 11.03.22; peer-reviewed by A Sikder, E Nabovati, Y Chu, G Strudwick; comments to author 20.06.22; revised version received 26.08.22; accepted 22.10.22; published 09.11.22.

Please cite as:
Riman KA, Davis BS, Seaman JB, Kahn JM
The Use of Electronic Health Record Metadata to Identify Nurse-Patient Assignments in the Intensive Care Unit: Algorithm Development and Validation
JMIR Med Inform 2022;10(11):e37923
URL: https://medinform.jmir.org/2022/11/e37923
doi:10.2196/37923
PMID:36350679

Original Paper

Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study

Dan Ouchi[1,2], MSc; Maria Giner-Soriano[1,2,3], PhD; Ainhoa Gómez-Lumbreras[1,4], PhD; Cristina Vedia Urgell[1,2,3], MD; Ferran Torres[5], PhD; Rosa Morros[1,2,3,6], PhD

[1]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina, Barcelona, Spain

[2]Facultat de Medicina. Departament de Farmacologia, Toxicologia i Terapèutica, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallés), Spain

[3]Unitat de Farmacia. Servei d'Atenció Primaria Barcelonès Nord i Maresme, Institut Català de la Salut, Barcelona, Spain

[4]Department of Pharmacotherapy, College of Pharmacy, University of Utah, Salt Lake City, UT, United States

[5]Unitat de Bioestadística Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Spain

[6]Spanish Clinical Research Network Platform, Barcelona, Spain

Corresponding Author:
Dan Ouchi, MSc
Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina
Gran Via de les Corts Catalanes 587, àtic
Barcelona, 08007
Spain
Phone: 34 934824110
Email: douchi@idiapjgol.info

Abstract

Background: Since the use of electronic health records (EHRs) in an automated way, pharmacovigilance or pharmacoepidemiology studies have been used to characterize the therapy using different algorithms. Although progress has been made in this area for monotherapy, with combinations of 2 or more drugs the challenge to characterize the treatment increases significantly, and more research is needed.

Objective: The goal of the research was to develop and describe a novel algorithm that automatically returns the most likely therapy of one drug or combinations of 2 or more drugs over time.

Methods: We used the Information System for Research in Primary Care as our reference EHR platform for the smooth algorithm development. The algorithm was inspired by statistical methods based on moving averages and depends on a parameter *Wt*, a flexible window that determines the level of smoothing. The effect of *Wt* was evaluated in a simulation study on the same data set with different window lengths. To understand the algorithm performance in a clinical or pharmacological perspective, we conducted a validation study. We designed 4 pharmacological scenarios and asked 4 independent professionals to compare a traditional method against the smooth algorithm. Data from the simulation and validation studies were then analyzed.

Results: The *Wt* parameter had an impact over the raw data. As we increased the window length, more patient were modified and the number of smoothed patients augmented, although we rarely observed changes of more than 5% of the total data. In the validation study, significant differences were obtained in the performance of the smooth algorithm over the traditional method. These differences were consistent across pharmacological scenarios.

Conclusions: The smooth algorithm is an automated approach that standardizes, simplifies, and improves data processing in drug exposition studies using EHRs. This algorithm can be generalized to almost any pharmacological medication and model the drug exposure to facilitate the detection of treatment switches, discontinuations, and terminations throughout the study period.

(JMIR Med Inform 2022;10(11):e37976) doi:10.2196/37976

KEYWORDS

electronic health records; data mining; complex drug patterns; algorithms; drug utilization; polypharmacy; EHR; medication; drug combination; therapy; automation; drug exposition; treatment; adherence

Introduction

The recent rise in the use of electronic health records (EHRs) has had a major impact on epidemiological research. These databases provide a low-cost means of accessing longitudinal data such as demographic, vital signs, administrative, medical and pharmacy claims, clinical, and patient-centered data on large populations for epidemiologic research [1,2].

However, in their current form, EHRs are complex and imperfect data sets that can be enhanced in a dizzying number of often ineffective ways. Although the challenges of working with EHRs in clinical trials have been identified [3-5], more research is needed to develop new and better ways to use them.

From the data mining perspective, addressing these data is particularly challenging as the outcomes can be significantly affected depending on the quality, validity, completeness, and heterogeneity of the available data [6]. Besides technical perspective, researchers have their particular ways of addressing EHR, dealing with EHR complexity, and their decisions have been shown to have a significant impact to the results [7,8]. Approaching these challenges in a heterogeneous and biased way favors the emergence of inconsistencies between similar studies [9].

In studies with EHR-based databases, information on drug exposure is usually obtained from electronic prescription, electronic dispensation, or invoice of drugs. This information is widely used and accepted in clinical research as the availability of longitudinally recorded data allows for a detailed characterization of both the exposure to medication and the outcome of interest, and mining the data contained within EHRs can potentially generate a greater understanding of medication effects in the real world, complementing what we know from randomized control trials [10].

Focusing on pharmacovigilance or pharmacoepidemiology when using EHRs, one of the main objectives is to characterize the therapy in terms of duration [11], discontinuation [12,13], changes [14], and adherence to pharmacological treatments [15]. Although progress has been made in this area for monotherapy [16], when we study treatment exposure in diseases such as hypertension, diabetes, or chronic obstructive pulmonary disease, treatment often switches from monotherapy to combinations of two or more drugs, which significantly increases the challenge of characterizing the treatment. In our experience [17], polytherapy in EHR-based studies creates complex treatment patterns that are challenging to analyze or interpret, can be blinded to researchers, and can be a source of misunderstanding as it is difficult to distinguish whether they are real occurrences or recording errors. To address this, we propose a novel algorithm called *smooth* to obtain the most likely therapy of one or more drugs over time.

Methods

Data Sources

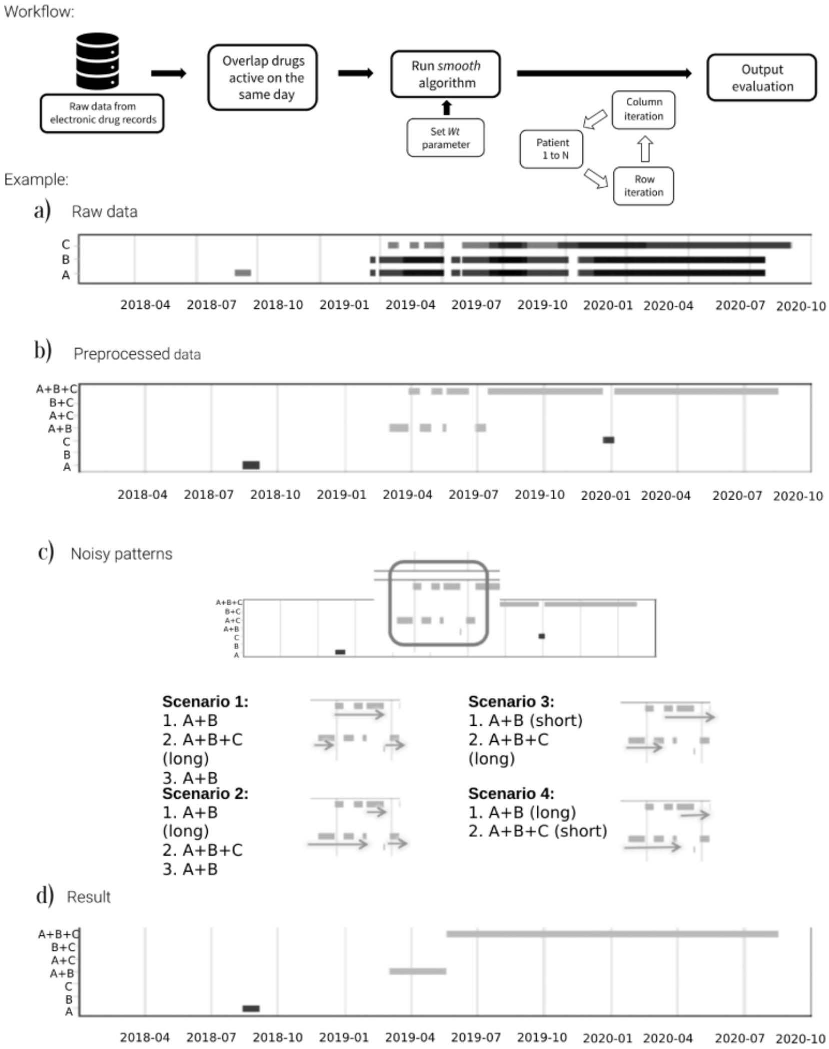
We used the Information System for Research in Primary Care (SIDIAP) [18] as our reference EHR platform for the algorithm development. The SIDIAP includes information recorded by health professionals during routine visits at 287 primary health care centers from the Catalan Health Institute (Institut Català de la Salut).

The platform includes information on disease diagnoses (*International Classification for Diseases, 10th Edition*), drug prescriptions and drug invoices in the primary care setting (Anatomical Therapeutic Chemical [ATC] classification system), and clinically relevant parameters (eg, weight, blood pressure, laboratory tests) as well as sociodemographic characteristics. It is also linked to a hospital discharge database for patients admitted to the Catalan Health Institute hospitals (30% of the SIDIAP population). The SIDIAP has pseudonymized records for more than seven million people and is representative of the Catalan population in terms of age, sex, and geographic distribution [19].

For the algorithm development and validation study, we used a subset of patients drawn from all Catalan Health Institute primary care centers. From this population, we also obtained sociodemographic characteristics: sex, age, country of origin, profession, socioeconomic index, smoking habits, alcohol intake, institutionalization in nursing homes, comorbidities, and electronic prescriptions of pharmacological treatments.

In SIDIAP, electronical prescriptions and drug invoices are stored in longitudinal format. Each record comprises the pseudonymized patient identifier, ATC code, and prescription or invoice date. The end of the prescription is determined by the health professional, whereas in drug invoice records we only have the month in which the invoice was made, and thus the end of the treatment is usually inferred based on the number of packages collected. Each prescription or invoice is recorded independently of the health problem, and duplicate records or overlaps are common (Figure 1A).

Figure 1. Overview of the smooth algorithm workflow with an illustrated example. (A) Horizontal axis is follow-up time in days, and the vertical axis is the different drugs prescribed. Box length indicates the period in which the prescription is active. (B) Patient profile after combining all active prescriptions in the same day, the first step of data process. (C) Example of complex patterns and 4 ways to overcome them. (D) Result obtained after passing the data through the smooth algorithm.



Ethics Approval

The study protocol was approved by the Research Ethics Committee of Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol; AR20/029, SIDIAP 386 on June 3, 2020). This is a database research study that has been conducted according to the guidelines of the Declaration of Helsinki (Fortaleza, Brazil 2013) and does not require consent from the people included to participate or for publication. The need for consent was waived by the Research Ethics Committee of IDIAPJGol as it is deemed unnecessary according to European legislation (Regulation [EU] 2016/679).

Finding the Most Likely Therapy Using the Smooth Algorithm

The process has 2 parts: data mining to look for treatments recorded daily and applying the smooth algorithm to the data (Figure 1).

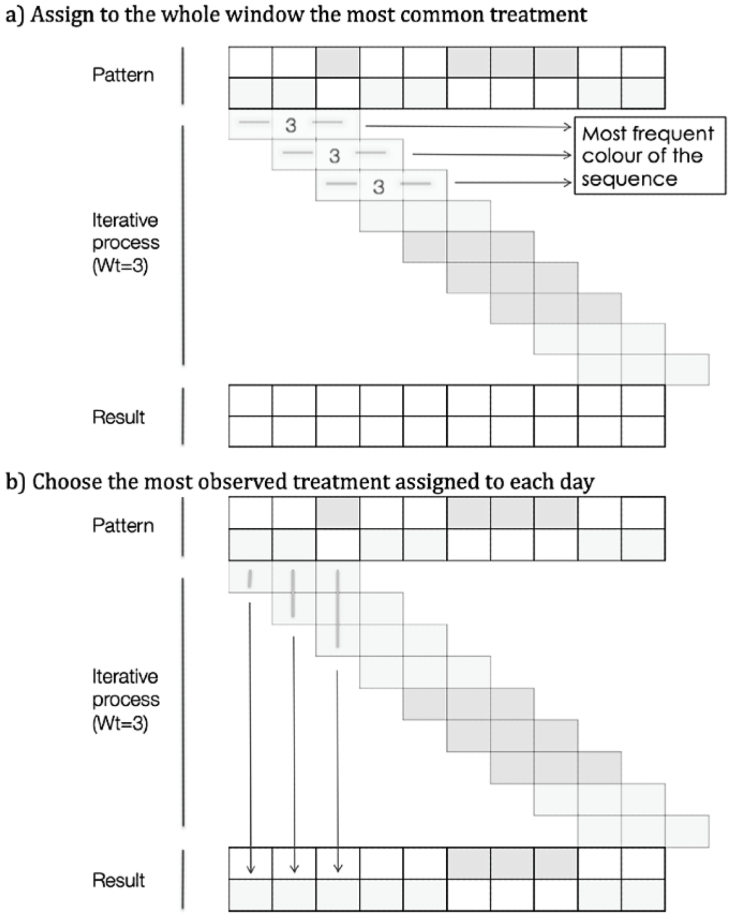
For the first part, we looked at all drugs of interest that are active on the same day (Figure 1A and Figure 1B). This step simplifies the prescription or invoice records but frequently reveals complex patterns that should be considered before conducting any analysis. Data are processed based on assumptions about those patterns by individual researchers and therefore give nonhomogeneous results. In Figure 1C, we imagined 4 scenarios (but there could be more) to handle the same problem (highlighted area during the first half of 2019). While some researchers may consider that the first therapy lasts until we observe a change in the treatment, others with different backgrounds may decide that a change from double to triple therapy can only be considered if the triple therapy lasts longer than an arbitrary period (eg, 60 days).

The smooth algorithm is inspired by statistical methods using simple moving averages that calculate trends or smooth time series [20]. For EHRs, we changed the concept of moving averages to a moving window from where we choose the most frequent treatment. Thus, by moving the window one day at a time, we identify the most frequent pattern over the study period (Figure 1D).

In Figure 2, we can see a detailed description of the algorithm. It is an iterative process that works as follows:

- Starting from the first prescription or dispensation at day t_i, we opened a window of specific length in days (*Wt*) in which we search for the most frequent treatment
- That treatment was assigned to the whole window unless we had a draw (2 treatments are active for the same number of days), in which case we carried the previous treatment (t_{i-1}) forward.
- We then shifted the window forward one day at a time, repeating the process until the end of follow-up (Figure 2A).
- After the first iteration, we had up to *Wt* possible treatments (or candidates) for each day
- Finally, we chose the candidate most frequently observed on that day (Figure 2B).

Figure 2. Smooth algorithm in detail.

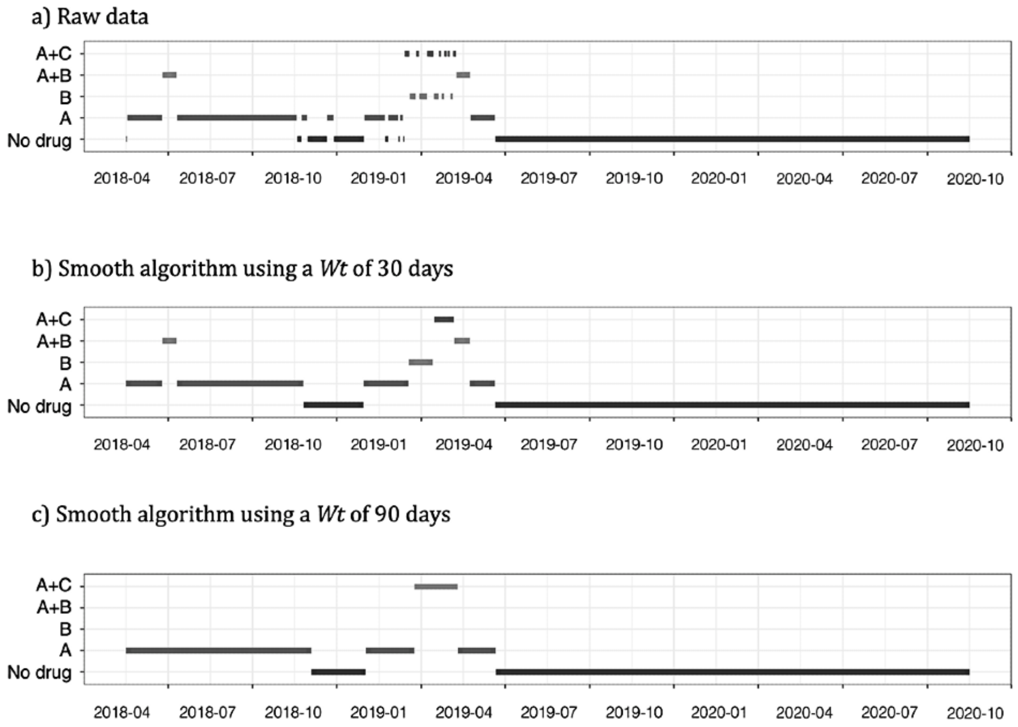

a) Assign to the whole window the most common treatment
Pattern
3
3
3
Most frequent colour of the sequence
Iterative process (Wt=3)
Result
b) Choose the most observed treatment assigned to each day
Pattern
Iterative process (Wt=3)
Result

Window Size

The length of *Wt* is the only parameter that needs to be defined beforehand, and its value can modify the outcome (Figure 3). The length of *Wt* determines the level of smoothing, and the value can go from 1 day to the total days of follow-up. Thus, for *Wt* of 1 day, we are not changing the data while for a *Wt* equal to the number of days of follow-up, we expect to reduce all records to the most frequent treatment. Therefore, small values of the *Wt* parameter will not significantly modify the raw data, whereas increasing the size of the window is expected to have a larger impact on the data.

In a more in-depth analysis, a simulation study was conducted using 7132 patients who were under long-term administration of aspirin, statins, beta-blockers, and angiotensin-converting enzyme (ACE) inhibitors or angiotensin-receptor blockers between 2018 and 2020. The smooth algorithm was run on this data using 6 *Wt* values: 10, 20, 30, 45, 60, and 90 days. For each *Wt*, we counted the number of patients with at least one change in the treatment pattern; out of these, we calculated the percentage of smooth as the ratio of the number of days changed by the algorithm divided by the total number of days with active treatment.

Figure 3. Example of how the outcome changes according to the value of the Wt parameter.



Validation Study

To understand how the algorithm performs from a clinical or pharmacological perspective, we conducted a validation study. We identified 4 pharmacological scenarios where the algorithm could be needed—combination of 3 or more drugs (platelet aggregation inhibitors, beta-blockers, and ACE inhibitors), treatments likely to discontinue (antidepressants), long-term combination of 2 drugs (insulins and oral antidiabetics), and short-term treatment (systemic antibiotics)—and asked to 4 independent professionals with experience in databases and drug exposure studies to compare a traditional method with the smooth algorithm.

The traditional method is a more intuitive and simple approach, commonly observed in the literature, to address noise and variability in electronic drug records [21,22]. Briefly, it starts with the first treatment observed and accepts a treatment change only if the new one is longer than a certain period of time. This period is generally arbitrary, an assumption done by the researcher based on the characteristics of the drug. For our validation study, we set the period to 60 days except for antibiotics (the short-term treatment), with a period of 15 days. For the smooth algorithm, the *Wt* parameter was set to 60 days in all 4 scenarios.

Before conducting the validation, we prepared a training session with the 4 reviewers consisting in an introduction to the data, drugs of study (including the selected ATC codes), explanation of the algorithms, and discussion of the common criteria to apply during the validation. From their feedback, and after the training, we decided to include all health problems related to the treatment as it may facilitate the evaluation and give more importance to clinical criteria (see validated sample in Multimedia Appendix 1, Figure S1).

Our primary objective was changes to the original data; we analyzed whether the algorithms improved, worsened, or made no changes to the original data. In addition, we asked the reviewers to choose between the traditional method and the smooth algorithm and evaluate its value for detecting treatment switches and/or discontinuations.

Each professional reviewed 100 patient records with one-quarter of the records being assigned to all reviewers to analyze consistency across validations. To reduce potential biases, reviewers were blinded and they did not know which method or algorithm generated the results (Table 1).

Table 1. Description of drugs and distribution of samples in the validation study.

Treatment pattern of use	Description (ATC[a] code)	Prescriptions in the data set	Samples analyzed, total (per reviewer)	Samples repeated across reviewers, total (per reviewer)
Short-term drugs	Systemic antibiotics (J01)	10,846,282	80 (20)	40 (10)
Likely to discontinue	Antidepressants (N06a)	3,859,496	80 (20)	40 (10)
Long-term combinations of 2 drugs	Insulins and oral antidiabetics (A10)	22,271,154	120 (30)	60 (15)
Combination of 3 or more drugs	Platelet aggregation inhibitors (B01Ac)	21,253,742	120 (30)	60 (15)

[a] ATC: Anatomical Therapeutic Chemical.

Statistical Methods

We determined that a sample size of 400 patients would be enough to ensure 80% power assuming a minimum effect size of 0.3062 with two degrees of freedom for a Chi-square test under a significance level of 5%.

Categorical variables were described with relative and absolute frequencies, and results from numerical variables were reported using means and standard deviations. For the validation study, we used the Chi-square test to evaluate differences between the traditional method and smooth algorithm on performance compared with the raw data. The algorithm was programmed in R (version 4.1.0, R Foundation for Statistical Computing), and all analyses were performed in R.

Results

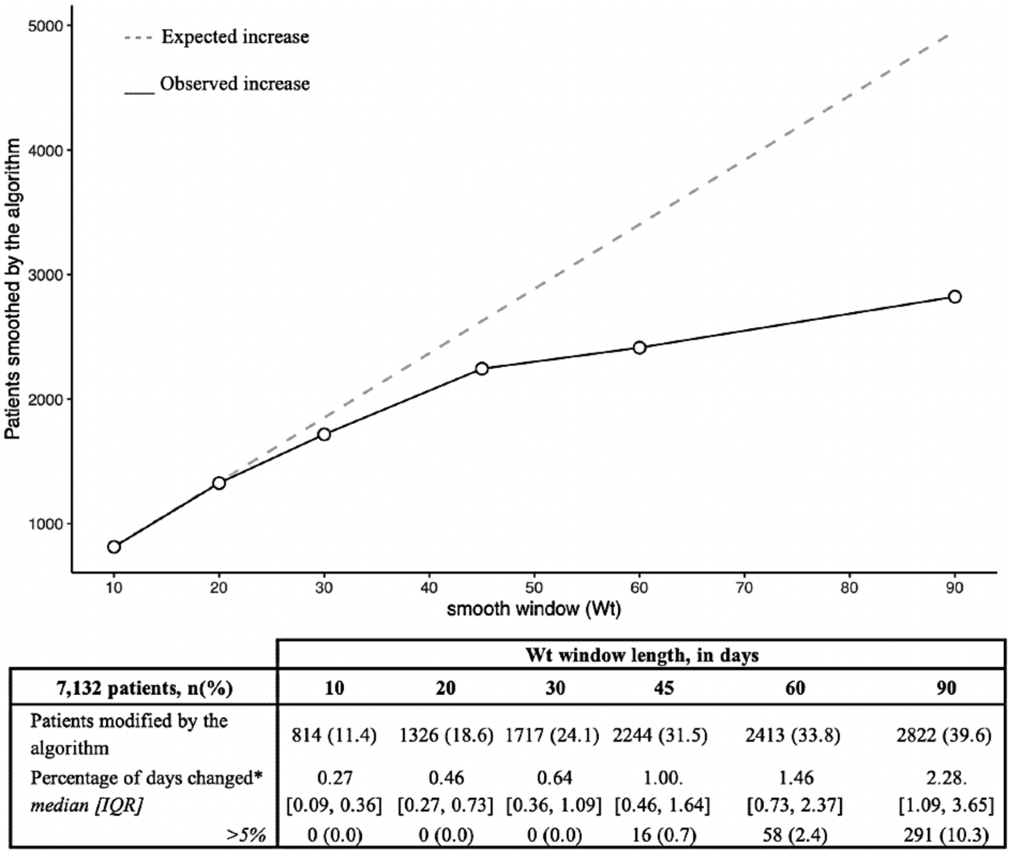
Impact of *Wt* Parameter and Simulation Study

In Figure 3, we show how the results changed according to the *Wt* value. In the example, for the raw data we note that during the first period of follow-up, the main treatment was a monotherapy followed by a complex pattern during the first 6-month semester of 2019 (Figure 3A). To reduce noise in the pattern, we applied the smooth algorithm using *Wt* values of 30 and 90 days. With a *Wt* of 30 days (Figure 3B), we retained the combination of 2 drugs during the early stages of follow-up; during months with more changes, patient moved from monotherapy to a combination of 2 drugs (A+C and A+B) prior to switching to the A monotherapy. In contrast, with a *Wt* of 90 days (Figure 3C), the entire treatment pattern was simplified. During the first year of follow-up, we observed a monotherapy; during the most complex pattern, the algorithm smoothed the changes to a single combination of A+C before returning to a monotherapy.

Results of the simulation study are represented in Figure 4. With a *Wt* of 10 days, 11.4% (814/7132) had their treatment pattern changed, while with a 90-day window, 39.6% (2822/7132) had their treatment pattern modified; 31.5% (2244/7132), 33.8% (2413/7132), and 39.6% (2822/7132) of patients with Wt values of 45, 60, and 90 days, respectively, saw at least 1 change. Thus, the effect of *Wt* was not linear as the expected progression of the number of patients being smoothed was different than the results from the simulation.

As a relative measure, we reported the percentage of days changed, ranging from 0.27 (IQR 0.09, 0.36) to 2.28 (IQR 1.09, 3.65). Thus, for each of the *Wt* values 10, 20, 30, 45, 60, and 90 days in a 1000-day period of follow-up, the algorithm modified 2.7, 4.6, 6.4, 10, 14.6, and 22.8 days, respectively. At windows from 10 to 30 days, the percentage of days changed always remain below the 5%, but as we increased the *Wt* to 45, 60, and 90 days, we started to observe patients with more than 5% of the data smoothed.

Figure 4. Statistics of the simulation study using 6 Wt values on 7132 patients under treatment for cardiovascular disease between 2018 and 2020. *100 x Number of days changed / Total days under prescription.



	Wt window length, in days					
7,132 patients, n(%)	10	20	30	45	60	90
Patients modified by the algorithm	814 (11.4)	1326 (18.6)	1717 (24.1)	2244 (31.5)	2413 (33.8)	2822 (39.6)
Percentage of days changed* *median [IQR]*	0.27 [0.09, 0.36]	0.46 [0.27, 0.73]	0.64 [0.36, 1.09]	1.00. [0.46, 1.64]	1.46 [0.73, 2.37]	2.28. [1.09, 3.65]
>5%	0 (0.0)	0 (0.0)	0 (0.0)	16 (0.7)	58 (2.4)	291 (10.3)

Validation Study

Multimedia Appendix 2 includes the results of the validation study. For the 400 samples, the smooth algorithm improved the raw data for 56.8% (227/400) of individuals, while 42.5% (170/400) benefited from using the traditional method. In 39% (156/400) of the samples, the outcome provided by each algorithm did not change the patterns, and 4.2% (17/400) of cases reported worsening after being processed by the smooth algorithm. The traditional method resulted in a worse outcome for 18.5% (74/400) of the samples, and the observed differences between algorithms were statistically significant (*P*<.001).

Significant differences were also observed between algorithms stratified by scenario. With a combination of 3 or more drugs (platelet aggregation inhibitors, beta-blocker, and ACE inhibitors), drugs likely to be discontinued (antidepressants), and the combination of 2 or more long-term drugs (insulins and antidiabetics), the smooth algorithm improved 69.2% (83/120), 61.3% (49/80), and 60.0% (72/120) of samples, respectively. In short-term treatment (systemic antibiotics), 90.0% (72/80) of samples did not show changes and 28.7% (23/80) were improved by the smooth algorithm.

As for the samples validated by the 4 professionals (Table 1), they decided in 88.0% (44/50) of patients to choose the smooth algorithm over the traditional method (Multimedia Appendix 2 and Multimedia Appendix 1, Figure S2). It was in the short-term treatment scenario where we observed less of a consensus, 70% (7/10), whereas for the rest of scenarios the 4 reviewers agreed in 93.3% (14/15), 100% (10/10), and 86.7% (13/15) of the samples, respectively.

The smooth algorithm performed better than the traditional method in detecting discontinuations (350/366, 95.6%) and treatment switches (138/230, 60.0%; see Multimedia Appendix 1, Table S1).

Discussion

Principal Findings

Since the use of EHR databases began in pharmacoepidemiologic studies, researchers have been trying to establish algorithms to model drug exposure [23]. This becomes even more challenging when trying to assess drug exposures with multiple pharmacologic treatments, which happens quite often in older people, as they are prescribed with up to 5 drugs simultaneously [24]. Thus, we have developed an automatic algorithm to model drug exposure through EHRs, which standardizes the data mining process to obtain more consistent and replicable results across studies.

The algorithm is inspired by time series forecasting methods and requires a parameter to be set beforehand. This is commonly observed in similar statistical methods such as autoregressive models or moving averages [25], and it is known that the value of the parameter can modify the outcome significantly [26-28]. The simulation study shows the impact of the *Wt* value. Small values hardly change the original data, but as the parameter value increases, the raw data can be affected to the point of losing clinical relevance. In the worst-case scenario, we observed changes in up to 40% of the patients, with 75% of those having at least 1% of the records smoothed.

Interestingly, the simulation shows that at a certain *Wt* value, the number of individuals modified reaches a plateau. The data changed by the algorithm are less than expected, particularly when *Wt* is greater than 30 days, suggesting that independently of the parameter, some patients will never be changed by the algorithm.

In the validation study, we observed that most times our algorithm improved the data patterns. It was designed to improve polypharmacy exposure assessment, and we were interested in the results for combinations of 3 or more drugs. In this scenario, both the traditional and smooth approaches demonstrated usefulness, and the percentages of improved samples were similar, although the smooth algorithm performed significantly better. These differences were also observed in the other scenarios, and we believe that the smooth algorithm not only improves the treatment pattern but also does not make it worse. In addition, the performance was not affected by the window length, since for antibiotics and antidepressants (short- and long-term drug use, respectively) the smooth algorithm performed well using the same *Wt* window.

The traditional method proved to be a good approach, and similar versions are being used in other studies [21,22,29], but it differs according to drug or study characteristics and so is less generalizable. In fact, it has not worked well for systemic antibiotics even though we specifically changed it to fit for its characteristics. Overall, the validations for the smooth algorithm were consistent between scenarios and reviewers.

Potential Uses and Strengths

We believe that the smooth algorithm has significant potential to assess exposure for treatment combinations, especially in chronic treatments, since it allows us to have a time sequence of exposure to the treatment. This sequence allows us to better model the drug exposition and detect discontinuations, switches, and periods of interaction with other drugs of short duration. In addition, it can be of great help in estimating adherence to a combination treatment [30]. With the smooth algorithm, we can easily calculate the exposure time using only electronic prescriptions without the need to know the dosage posology.

From a clinical point of view, the smooth algorithm has great advantage when estimating polypharmacy adherence [31]. Patients affected by chronic conditions in need of polypharmacy may have differing levels of adherence to individual medications within their regimen, and this could lead to varying health outcomes and misleading results if the methodological approach assumes as equivalent adherence to all medications. These patients also face other acute conditions requiring the addition of drugs or modification of doses while maintaining their actual medication regimen.

Another research area in which our algorithm could be of utility is the study of adverse events due to lack of effectivity (antibiotics or hypertension treatments) and drug-drug interactions (anticoagulants and nonsteroidal anti-inflammatories combined for short time periods).

Although we cannot ever know a patient's true adherence, the smooth algorithm is an automated way to analyze EHR data offering methodological consistency across studies. In several studies, assumptions were made prior to the data processing, and this may have impacted the results of the analyses by introducing bias on the final results [32]. Algorithms like smooth can help standardize these assumptions and minimize inconsistencies between studies with similar databases.

Limitations

Due to the nature of the algorithm eliminating complex patterns, we may lose relevant exposures of a short period of time, and smooth is not recommended for all scenarios. Similarly, smooth may not work well in long-acting drugs.

We were not able to capture the posology with the SIDIAP database, so we could not estimate the length of treatment through the dose prescribed. The treatment doses can change throughout the year (decreased use of diuretics during summertime, when traveling, etc).

From a technical perspective, using the smooth algorithm is a time-consuming process. To run the algorithm on big data sets like EHRs, good information technology is needed. The time needed to complete the process may vary depending on the number of patients and follow-up time.

In addition, before running the algorithm, we must set a parameter, *Wt*, that allows us to choose between precision and simplicity. Setting this parameter is not straightforward, and it is important to understand the effect on the outcome to use a good value. This is an inherent limitation of the algorithm and, as a guide, we recommend setting the *Wt* value within the range of 30 to 60 days to reduce complex patterns without compromising relevant information. Moreover, the *Wt* can be changed so the algorithm can be used in short (antibiotics) and long-term (antidepressants) treatments as well as in drug combinations for chronic conditions such as diabetes and hypertension.

Another limitation of the study is that the validation was done with the traditional method instead of other algorithms as a comparator. Moreover, in our experience, we commonly see the traditional method being used with some differences or criteria according to the framework or objectives of the study. For example, in projects with the European Medicines Agency, we have never seen a method or an automatic approach to deal with drug exposure other than the one we call traditional [33].

Conclusion

The smooth algorithm is an automated approach to estimate the most likely drug exposure pattern. We proved that it standardizes, simplifies, and improves the data processing steps before performing the study analysis; can model the drug exposure to detect cotreatment, switches, discontinuations, and treatment terminations; and facilitates adherence calculations throughout the study period. In future pharmacoepidemiological studies, we aim to further validate the algorithm and analyze the impact the algorithm can have on the main results.

Authors' Contributions

DO designed and developed the algorithm and performed the statistical analyses. MG and DO collected the data. All authors contributed to the validation study. DO, AG, MG, FT, and RM were involved in the interpretation of the data. DO wrote the manuscript and designed the tables and figures. All authors critically revised the report and approved the final version to be submitted for publication. The corresponding author confirms that he had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File , 1022 KB - medinform_v10i11e37976_app1.docx]

Multimedia Appendix 2
Results of the validation study.
[DOCX File , 15 KB - medinform_v10i11e37976_app2.docx]

References

1. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016 Mar;37:61-81. [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
2. Rogers JR, Lee J, Zhou Z, Cheung YK, Hripcsak G, Weng C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. J Am Med Inform Assoc 2021 Jan 15;28(1):144-154. [doi: 10.1093/jamia/ocaa224] [Medline: 33164065]
3. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017 Jan;106(1):1-9 [FREE Full text] [doi: 10.1007/s00392-016-1025-6] [Medline: 27557678]

4. Lu Z. Information technology in pharmacovigilance: benefits, challenges, and future directions from industry perspectives. Drug Healthc Patient Saf 2009 Oct;1:35-45 [FREE Full text] [doi: 10.2147/dhps.s7180] [Medline: 21701609]
5. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011 May;4:47-55 [FREE Full text] [doi: 10.2147/RMHP.S12985] [Medline: 22312227]
6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
7. Salas M, Lopes LC, Godman B, Truter I, Hartzema AG, Wettermark B, et al. Challenges facing drug utilization research in the Latin American region. Pharmacoepidemiol Drug Saf 2020 Nov 17;29(11):1353-1363. [doi: 10.1002/pds.4989] [Medline: 32419226]
8. Rudin RS, Friedberg MW, Shekelle P, Shah N, Bates DW. Getting value from electronic health records: research needed to improve practice. Ann Intern Med 2020 Jun 02;172(11_Supplement):S130-S136. [doi: 10.7326/m19-0878]
9. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, Joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. Pharmacoepidemiol Drug Saf 2017 Sep 15;26(9):1018-1032 [FREE Full text] [doi: 10.1002/pds.4295] [Medline: 28913963]
10. Yao L, Zhang Y, Li Y, Sanseau P, Agarwal P. Electronic health records: implications for drug discovery. Drug Discov Today 2011 Jul;16(13-14):594-599. [doi: 10.1016/j.drudis.2011.05.009] [Medline: 21624499]
11. Støvring H, Pottegård A, Hallas J. Determining prescription durations based on the parametric waiting time distribution. Pharmacoepidemiol Drug Saf 2016 Dec 26;25(12):1451-1459. [doi: 10.1002/pds.4114] [Medline: 27670969]
12. Lyu H, Zhao S, Yoshida K, Tedeschi S, Xu C, Nigwekar S, et al. Delayed denosumab injections and bone mineral density response: an electronic health record-based study. J Clin Endocrinol Metab 2020 May 01;105(5):1 [FREE Full text] [doi: 10.1210/clinem/dgz321] [Medline: 31894244]
13. Mascarenhas J, Mehra M, He J, Potluri R, Loefgren C. Patient characteristics and outcomes after ruxolitinib discontinuation in patients with myelofibrosis. J Med Econ 2020 Jul 31;23(7):721-727 [FREE Full text] [doi: 10.1080/13696998.2020.1741381] [Medline: 32159402]
14. Sharman J, Kabadi SM, Clark J, Andorsky D. Treatment patterns and outcomes among mantle cell lymphoma patients treated with ibrutinib in the United States: a retrospective electronic medical record database and chart review study. Br J Haematol 2021 Feb 23;192(4):737-746. [doi: 10.1111/bjh.16922] [Medline: 33095453]
15. Li X, Cole SR, Westreich D, Brookhart MA. Primary non-adherence and the new-user design. Pharmacoepidemiol Drug Saf 2018 Apr 19;27(4):361-364 [FREE Full text] [doi: 10.1002/pds.4403] [Medline: 29460385]
16. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet 2019 Nov;394(10211):1816-1826. [doi: 10.1016/s0140-6736(19)32317-7]
17. Giner-Soriano M, Sotorra Figuerola G, Cortés J, Pera Pujadas H, Garcia-Sangenis A, Morros R. Impact of medication adherence on mortality and cardiovascular morbidity: protocol for a population-based cohort study. JMIR Res Protoc 2018 Mar 09;7(3):e73 [FREE Full text] [doi: 10.2196/resprot.8121] [Medline: 29523501]
18. Bolíbar B, Fina AF, Morros R, Garcia-Gil MDM, Hermosilla E, Ramos R, et al. [SIDIAP database: electronic clinical records in primary care as a source of information for epidemiologic research]. Med Clin (Barc) 2012 May 19;138(14):617-621. [doi: 10.1016/j.medcli.2012.01.020] [Medline: 22444996]
19. García-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, Ramos R, et al. Construction and validation of a scoring system for the selection of high-quality data in a Spanish population primary care database (SIDIAP). Inform Prim Care 2011 Jun 01;19(3):135-145 [FREE Full text] [doi: 10.14236/jhi.v19i3.806] [Medline: 22688222]
20. Watson GS. Smooth regression analysis. Sankhyā Indian J Stat 1964;Series A(26):359-372.
21. Xu H, Aldrich M, Chen Q, Liu H, Peterson N, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc 2015 Jan;22(1):179-191 [FREE Full text] [doi: 10.1136/amiajnl-2014-002649] [Medline: 25053577]
22. Hughey J, Colby J. Discovering cross-reactivity in urine drug screening immunoassays through large-scale analysis of electronic health records. Clin Chem 2019 Dec;65(12):1522-1531 [FREE Full text] [doi: 10.1373/clinchem.2019.305409] [Medline: 31578215]
23. Andrade SE, Kahler KH, Frech F, Chan KA. Methods for evaluation of medication adherence and persistence using automated databases. Pharmacoepidemiol Drug Saf 2006 Aug;15(8):565-574. [doi: 10.1002/pds.1230] [Medline: 16514590]
24. Pazan F, Wehling M. Polypharmacy in older adults: a narrative review of definitions, epidemiology and consequences. Eur Geriatr Med 2021 Jun 10;12(3):443-452 [FREE Full text] [doi: 10.1007/s41999-021-00479-3] [Medline: 33694123]
25. Hannan EJ, Rissanen J. Recursive estimation of mixed autoregressive-moving average order. Biometrika 1983;70(1):303. [doi: 10.1093/biomet/70.1.303-b]
26. Durbin J. Efficient estimation of parameters in moving-average models. Biometrika 1959 Dec;46(3/4):306. [doi: 10.2307/2333528]
27. Sandgren N, Stoica P, Babu P. On moving average parameter estimation. Eur Signal Process Conf 2012:2348-2351.

28. Żebrowska M, Dzwiniel P, Waleszczyk WJ. Removal of the sinusoidal transorbital alternating current stimulation artifact from simultaneous EEG recordings: effects of simple moving average parameters. Front Neurosci 2020 Jul 29;14:735 [FREE Full text] [doi: 10.3389/fnins.2020.00735] [Medline: 32848538]
29. van Staa T, Abenhaim L, Leufkens H. A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. J Clin Epidemiol 1994 Feb;47(2):183-189. [doi: 10.1016/0895-4356(94)90023-x]
30. Vlacho B, Mata-Cases M, Mundet-Tudurí X, Vallès-Callol J, Real J, Farre M, et al. Analysis of the adherence and safety of second oral glucose-lowering therapy in routine practice from the mediterranean area: a retrospective cohort study. Front Endocrinol (Lausanne) 2021 Jul 14;12:708372 [FREE Full text] [doi: 10.3389/fendo.2021.708372] [Medline: 34335477]
31. Franklin JM, Gopalakrishnan C, Krumme AA, Singh K, Rogers JR, Kimura J, et al. The relative benefits of claims and electronic health record data for predicting medication adherence trajectory. Am Heart J 2018 Mar;197:153-162. [doi: 10.1016/j.ahj.2017.09.019] [Medline: 29447776]
32. Pye SR, Sheppard T, Joseph RM, Lunt M, Girard N, Haas JS, et al. Assumptions made when preparing drug exposure data for analysis have an impact on results: an unreported step in pharmacoepidemiology studies. Pharmacoepidemiol Drug Saf 2018 Jul 17;27(7):781-788 [FREE Full text] [doi: 10.1002/pds.4440] [Medline: 29667263]
33. Cid-Ruzafa J, Lacy BE, Schultze A, Duong M, Lu Y, Raluy-Callado M, et al. Linaclotide utilization and potential for off-label use and misuse in three European countries. Therap Adv Gastroenterol 2022 Jun 11;15:17562848221100946 [FREE Full text] [doi: 10.1177/17562848221100946] [Medline: 35706826]

Abbreviations

ACE: angiotensin-converting enzyme
ATC: Anatomical Therapeutic Chemical
EHR: electronic health record
IDIAPJGol: Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina
SIDIAP: Information System for Research in Primary Care

Edited by C Lovis; submitted 14.03.22; peer-reviewed by E Urtnasan, G Yu; comments to author 19.08.22; revised version received 19.09.22; accepted 13.10.22; published 15.11.22.

Please cite as:
Ouchi D, Giner-Soriano M, Gómez-Lumbreras A, Vedia Urgell C, Torres F, Morros R
Automatic Estimation of the Most Likely Drug Combination in Electronic Health Records Using the Smooth Algorithm: Development and Validation Study
JMIR Med Inform 2022;10(11):e37976
URL: https://medinform.jmir.org/2022/11/e37976
doi:10.2196/37976
PMID:36378514

Original Paper

Linking Biomedical Data Warehouse Records With the National Mortality Database in France: Large-scale Matching Algorithm

Vianney Guardiolle[1], MSc, MD; Adrien Bazoge[1,2], MSc; Emmanuel Morin[2], PhD; Béatrice Daille[2], PhD; Delphine Toublant[1], MS; Guillaume Bouzillé[3], MD, PhD; Youenn Merel[3], MSc; Morgane Pierre-Jean[3], PhD; Alexandre Filiot[4], MSc; Marc Cuggia[3], MD, PhD; Matthieu Wargny[1], MSc, MD; Antoine Lamer[5], PhD; Pierre-Antoine Gourraud[1,6], PhD

[1]CHU de Nantes, INSERM CIC 1413, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, 44000, Nantes, France

[2]LS2N UMR CNRS 6004, Université de Nantes - 2, rue de la Houssinière - BP 92208 - 44322 Nantes Cedex 03 - France, Nantes, France

[3]Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099,35000, Rennes, France

[4]CHU Lille, INCLUDE: Integration Center of the Lille University Hospital for Data Exploration, 59000, Lille, France

[5]Univ. Lille, CHU Lille, ULR 2694, METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000, Lille, France

[6]Université de Nantes, CHU de Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ATIP-Avenir, Nantes, France

Corresponding Author:
Antoine Lamer, PhD
Univ. Lille, CHU Lille, ULR 2694
METRICS: Évaluation des Technologies de santé et des Pratiques médicales, F-59000
1 place de Verdun
Lille, 59000
France
Phone: 33 320626969
Email: antoine.lamer@univ-lille.fr

Abstract

Background: Often missing from or uncertain in a biomedical data warehouse (BDW), vital status after discharge is central to the value of a BDW in medical research. The French National Mortality Database (FNMD) offers open-source nominative records of every death. Matching large-scale BDWs records with the FNMD combines multiple challenges: absence of unique common identifiers between the 2 databases, names changing over life, clerical errors, and the exponential growth of the number of comparisons to compute.

Objective: We aimed to develop a new algorithm for matching BDW records to the FNMD and evaluated its performance.

Methods: We developed a deterministic algorithm based on advanced data cleaning and knowledge of the naming system and the Damerau-Levenshtein distance (DLD). The algorithm’s performance was independently assessed using BDW data of 3 university hospitals: Lille, Nantes, and Rennes. Specificity was evaluated with living patients on January 1, 2016 (ie, patients with at least 1 hospital encounter before and after this date). Sensitivity was evaluated with patients recorded as deceased between January 1, 2001, and December 31, 2020. The DLD-based algorithm was compared to a direct matching algorithm with minimal data cleaning as a reference.

Results: All centers combined, sensitivity was 11% higher for the DLD-based algorithm (93.3%, 95% CI 92.8-93.9) than for the direct algorithm (82.7%, 95% CI 81.8-83.6; *P*<.001). Sensitivity was superior for men at 2 centers (Nantes: 87%, 95% CI 85.1-89 vs 83.6%, 95% CI 81.4-85.8; *P*=.006; Rennes: 98.6%, 95% CI 98.1-99.2 vs 96%, 95% CI 94.9-97.1; *P*<.001) and for patients born in France at all centers (Nantes: 85.8%, 95% CI 84.3-87.3 vs 74.9%, 95% CI 72.8-77.0; *P*<.001). The DLD-based algorithm revealed significant differences in sensitivity among centers (Nantes, 85.3% vs Lille and Rennes, 97.3%, *P*<.001). Specificity was >98% in all subgroups. Our algorithm matched tens of millions of death records from BDWs, with parallel computing capabilities and low RAM requirements. We used the Inseehop open-source R script for this measurement.

Conclusions: Overall, sensitivity/recall was 11% higher using the DLD-based algorithm than that using the direct algorithm. This shows the importance of advanced data cleaning and knowledge of a naming system through DLD use. Statistically significant differences in sensitivity between groups could be found and must be considered when performing an analysis to avoid differential biases. Our algorithm, originally conceived for linking a BDW with the FNMD, can be used to match any large-scale databases. While matching operations using names are considered sensitive computational operations, the Inseehop package released here is easy to run on premises, thereby facilitating compliance with cybersecurity local framework. The use of an advanced deterministic matching algorithm such as the DLD-based algorithm is an insightful example of combining open-source external data to improve the usage value of BDWs.

(JMIR Med Inform 2022;10(11):e36711) doi:10.2196/36711

KEYWORDS

data warehousing; clinical data warehouse; medical informatics applications; medical record linkage; French National Mortality Database; data reuse; open data, R; clinical informatics

Introduction

Vital status is important information for medical research. While real-world evidence from the analysis of biomedical data warehouse (BDW) records has gained popularity in recent years [1], the longitudinal value of the information is weakened by the uncertainty of patients' vital statuses. This information is often limited to inpatients who died while hospitalized.

France has a long tradition of administrative centralization inherited from the Napoleonian era. When a resident dies on French territory, French city halls are required to send a death report to the Institut National de la Statistique et des Études Économiques (INSEE), translated to the French National Institute of Statistics and Economic Studies [2]. This report is used to complete the French National Mortality Database (FNMD). This database, which contains tens of millions of records, is updated monthly and has been open access since 2019 [3].

Record linkage (also referred to as data matching or entity resolution) is the process of quickly and accurately identifying records corresponding to the same individual entity from one or more data sources [4]. A recent literature review by Bounebache et al [5] presents record linkage and its multiple challenges. Two different approaches exist: (1) deterministic record linkage, which uses expert knowledge and possible statistical learning [5]; and (2) probabilistic linkage, which relies on a statistical model to evaluate the contribution of each variable in the record linkage strategy [6,7]. Matching large-scale BDW records with the FNMD presents multiple challenges. The first is the absence of a unique common identifier, such as a social security number. Second, surnames are shared within families and may change over one's lifetime based on varying cultural practices regarding marriage. Additionally, first and middle names can be confounded or compound. Third, clerical errors can occur when identities are administratively recorded in both databases [4]. Furthermore, in practice, the exponential number of comparisons often prohibits direct matching of millions of database records to the FNMD, which contains tens of millions of records.

Consequently, little has been published on the computational performance of record linkage, its accuracy, and its determinants. Moore et al [8] state that record-linkage performance must be evaluated to validate statistical analyses. For example, they showed that a specificity of <95% prevents estimating a significant risk ratio of 2. Previous studies [9,10] have used record linkage with the FNMD. Most of these studies used suboptimal references to evaluate algorithm performance, small databases (ie, <20,000 patients), were monocentric, or did not share their source code. Bannay et al [11] linked a BDW with Système National des Données de Santé (SNDS), translated to the French National Health Database. As the extraction from the SNDS was anonymized in accordance with national legislation, they implemented a semideterministic record linkage procedure based on the variables of the hospital discharge report (ie, sex, year of birth, month of birth, admission and discharge dates, diagnoses, etc).

To routinely update vital status in BDW records from the FNMD on a large scale, we developed a deterministic matching algorithm based on Damerau-Levenshtein distance (DLD) and compared its performance with that of a direct-matching algorithm as a reference for 3 regional hospital BDWs.

Methods

Data and Databases

FNDM files were downloaded from the national open data website [12] and included the following fields: birth surname, first name, middle names, birth date, sex, city and country of birth, death date, and zip code of the place of death. We found 11,490,867 records for the period between 2001 and 2020.

Three university hospitals in France were involved in this study: Lille, Nantes, and Rennes. Each hospital's BDW contains administrative, clinical, biological, and drug data.

In the Lille BDW, vital status information was available on June 1, 2021, for 1,609,515 patients who had at least 1 hospital stay between January 1, 2008, and June 1, 2021. The data showed that 1,570,320 (98%) patients were living, and 39,195 (2%) were deceased.

For the Nantes BDW, vital status information was available on January 14, 2021, for 2,035,805 patients who had at least 1 hospital encounter during the previous 20 years. The data showed that 1,974,786 (97%) patients were living and 61,019 (3%) were deceased.

For the Rennes BDW, vital status information was available on January 4, 2021, for 1,262,072 patients. The data showed that 1,221,817 (97%) patients were living, 37,986 (3%) were deceased, and 346 had no recorded vital status.

Hereafter, samples extracted from the BDWs are referred to as "local databases."

Record-Linkage Algorithms

To assess the performance gain induced by advanced data cleaning and DLD use, we used a simple direct-matching algorithm as a reference. Characteristics of the data cleaning and the algorithms used are presented in Multimedia Appendix 1.

Direct-Matching Algorithm as a Reference

The direct-matching algorithm removed accents from patients' first name and surname because the FNMD does not use accents. All letters were transformed into lowercase. Two records were linked between the local database and the FNMD if both records had the exact same surname, first name, birth date, and sex. The surname chosen was the birth surname if present or the current surname if not.

DLD-Based Algorithm as a Deterministic Solution

The DLD between 2 strings is the sum of necessary operations to transform string 1 into string 2, among insertion of a character, deletion of 1 character, substitution of 1 character by another, or transposition of 2 adjacent characters [13]. Examples are available in Multimedia Appendix 2.

Distances were calculated between the local database and FNMD for first name, surname, birth date, sex, and city of birth. For sex, a distance of 0 indicates that the sex is the same in both records, and a distance of 1 indicates a mismatch between the 2 records.

A 4-Step Algorithm

The algorithm can be divided into four consecutive steps: cleaning the data, creating new variables, validating pairs with blocking techniques, and choosing the more pertinent pairs.

Data Cleaning

In the FNMD, birth dates may be expressed with a missing day and month (eg, 1956-00-00). In these cases, the algorithm automatically attributed January 1 to the date to obtain a valid date format. If the birth date was invalid, the month and day were inverted and tested before choosing January 1. For example, the date 1960-31-03 is invalid (ie, there are not 31 months in a year); however, the date 1960-03-31 is valid, so 1960-03-31 was chosen. Another example is the date 1959-32-33: the dates 1959-32-33 and 1959-33-32 are also invalid; thus, the date 1959-01-01 was chosen.

Characters other than letters (eg, numbers and special characters) were removed from the local database and FNMD. All letters were changed to lowercase, and accents were removed.

In both the local database and FNMD, mentions of the district were suppressed, and only the city of birth was used. For example, "Paris, 13ème arrondissement" was changed "paris."

New Variable Creation

In the local database, a transformed city of birth variable was created wherein abbreviations were transformed into full text. For example, "St-Martin-sr-Ocre" was transformed into "saintmartinsurocre."

In the FNMD, the variable fnmd_firstname_0 was created from the first element of the first name (eg, "pierre" from "Pierre-Olivier"), and the variable fnmd_firstname_12 was created from concatenation of the first name and middle name (eg, "marieclaire" from first name "Marie" and middle name "Claire"). Other examples are available in Table 1.

Table 1. Examples of first name–related data created for first name Damerau-Levenshtein distance (DLD) computation.

Original data from the FNMD[a]		Data created for first name DLD[b] computation		
First name	First middle ame	fnmd_firstname_0	fnmd_firstname _1	fnmd_firstname_12
Jean	N/A[c]	jean	Jean	jean
Marie	Claire	marie	marie	marieclaire
Pierre-Olivier	Christian	pierre	pierreolivier	pierreolivierchristian
Elon-Louis	N/A	elon	elonlouis	elonlouis

[a] French National Mortality Database.

[b] Damerau-Levenshtein distance.

[c] Not applicable.

Pair Validation

Records from the local database and the FNMD matched if all the following conditions were valid: (1) the DLD of the first name was ≤ the maximal first name DLD, (2) the DLD of the surname was ≤ the maximal surname DLD, (3) the DLD of the birth date was ≤ the maximal birth date DLD, (4) the DLD of the sex was ≤ the maximal sex DLD, and (4) the total sum of the 4 previous DLDs (ie, the total DLD) was ≤ the maximal of the total DLD.

The DLD chosen for the surname was the shorter DLD among (1) the birth surname in the local database and the surname in the FNMD and (2) the current surname in the local database and the surname in the FNMD.

The DLD chosen for the first name was the shorter DLD among (1) the first name in the local database and fnmd_firstname_0, (2) the first name in the local database and fnmd_firstname _1, and (3) the first name in the local database and fnmd_firstname_12.

The DLD chosen for the birth city name (option for the more pertinent pairs selection) was the shorter DLD among (1) the original city of birth in the local database and city of birth in the FNMD and (2) the transformed city of birth in local database and city of birth in the FNMD.

Matching 2 databases (A and B) without a common identifier implies evaluating the match or nonmatch status of every element of AxB, called a pair [5]. The number of pairs to compare is given by the number of records in A multiplied by the number of records in B. This formula is particularly concerning when matching BDWs with the FNMDs, both of which potentially contain tens of millions of records, potentially leading to quadrillions of pairs to compare. Blocking techniques reduce the number of pairs to compare [5] and thus the execution time and RAM requirements. We successively applied a simple blocking technic 2 times, which consist of only comparing the pairs that contained the same value for 1 defined variable: first, the birth date and second, the concatenation of the first 4 characters of the first name and the first 4 characters of the surname (the birth surname when present and the current surname elsewhere). This allowed us to process the pairwise comparison, even if the birth date or the first 4 characters of first name/surname contained mismatches (but not if both birth date and the first 4 characters of the first name/surname contained mismatches).

Table 2 presents examples of pairs going through comparison process or not, given the 2 successive blocking processes used in the DLD-based matching algorithm.

Table 2. Examples of the blocking process for the Damerau-Levenshtein distance (DLD)–based matching algorithm.

FNMD[a] birth date	FNMD first name/surname concatenation[b]	Local database birth date	Local database first name/surname concatenation[b]	Comparison during birth date blocking	Comparison during first name/family name concatenation blocking
1935-06-29	louidefu	1935-06-29	louidefu	Yes	Yes
1935-06-29	louidefu	1931-10-08	maricall	No	No
1935-06-29	louidefu	1940-26-11	jeanpoku	No	No
1935-06-29	louidefu	1956-23-12	chardegu	No	No
1956-12-18	maricall	1935-06-29	louidefu	No	No
1956-12-18	maricall	1931-10-08	maricall	No	Yes
1956-12-18	maricall	1940-26-11	jeanpoku	No	No
1956-12-18	maricall	1956-23-12	chardegu	No	No
1940-11-26	jeanpoqu	1935-06-29	louidefu	No	No
1940-11-26	jeanpoqu	1931-10-08	maricall	No	No
1940-11-26	jeanpoqu	1940-11-26	jeanpoku	Yes	No
1940-11-26	jeanpoqu	1956-23-12	chardegu	No	No
1940-11-26	maricuri	1935-06-29	louidefu	No	No
1940-11-26	maricuri	1931-10-08	maricall	No	No
1940-11-26	maricuri	1956-23-12	chardegu	No	No
1940-11-26	maricuri	1956-23-12	chardegu	No	No

[a] FNMD: French National Mortality Database.

[b] Concatenation of the 4 first characters of the first name and the 4 first characters of the surname (birth surname if present, current surname elsewhere).

Choice of More Pertinent Pairs

One patient from a local database could be matched with none, 1, or multiple records from the FNMD. An algorithm is proposed in Multimedia Appendix 3 to select the most pertinent pairs for the last cases.

Data Sampling for Statistical Learning, Performance Evaluation, and Large-scale Testing

For all individuals, the following variables were extracted: birth surname, current surname, first name, birth date, sex, city and country of birth, vital status, and, if present, death date.

To learn the optimal parameters of the DLD-based algorithm, we randomly selected 3600 patients from the Nantes BDW, with 450 per stratum: (1) men born in France (MBIF) who died between 2001 and 2020 (deceased MBIF). These were the MBIF whose deaths were registered in the BDW between 2001 and 2020; (2) MBIF alive on 1 January 2016 (living MBIF). These were the MBIF with at least 1 hospital encounter between January 1, 2001, and December 31, 2015, and another hospital encounter between January 2, 2016, and December 31, 2020; (3) women born in France (WBIF) who died during between 2001 and 2020 (deceased WBIF); (4) WBIF who were alive on January 1, 2016 (living WBIF); (5) men born outside France (MBOF) who died between 2001 and 2020 (deceased MBOF); (6) MBOF who were alive on January 1, 2016 (living MBOF); women born outside France (WBOF) who died between 2001 and 2020 (deceased WBOF); and (7) WBOF who were alive on January 1, 2016 (living WBOF).

For the DLD-based algorithm, a maximal DLD of 2 was learned for the first name, 1 for the surname, 1 for the birth date, 1 for sex, and 2 for the total DLD.

To evaluate the specificity and sensitivity of both the DLD-based and direct algorithms, samples of 8000 patients were randomly extracted from each of the 3 BDWs. The sample from Nantes did not contain patients used for statistical learning of the DLD-based algorithm parameters (ie, the maximal DLDs). Each sample contained 1000 deceased MBIF, 1000 living MBIF, 1000 deceased WBIF, 1000 living WBIF, 1000 deceased MBOF, 1000 living MBOF, 1000 deceased WBOF, and 1000 living WBOF.

Finally, to assess the low RAM requirements and parallel processing capabilities of the DLD-based algorithm, a sample of 2 million patients was randomly extracted from the Nantes BDW, and 11 million records (between 2001 and 2020) were randomly extracted from the FNMD.

Specificity and Sensitivity Evaluated on Separate Data Sets

Sensitivity and specificity were evaluated for 8000 patients from every hospital for both the direct and DLD-based algorithms. Sensitivity (or recall) was evaluated for patients who died between January 1, 2001, and December 31, 2020, and were registered as the gold standard for each BDW. The algorithm classified patients as deceased if they were linked by at least one FNMD record. Specificity was evaluated for patients alive on 1 January 2016 (ie, patients with at least 1 hospital encounter between January 1, 2001, and December 31, 2015, and another hospital encounter between January 2, 2016, and December 31, 2020). The algorithm classified a patient as alive if the patient was not linked to the FNMD. The same patient could be present in both the sensitivity and specificity data sets (eg, a patient who died on May 13, 2017). This was not a problem because to evaluate specificity, we only used deaths registered in the FNMD between January 1, 2001, and January 1, 2016. To evaluate sensitivity, we used all deaths registered in the FNMD between January 1, 2001, and December 31, 2020. Specificity and sensitivity were calculated for each maximal total distance parameter of the DLD-based algorithm, from 0 to 5.

Our gold standard for the matching algorithms was for it to be reliable both for sensitivity and specificity evaluation. First, it was completely independent from the FNMD. Second, deceased status in the hospital databases was reliable because vital status at discharge is a necessary information for the stay fee payment to the hospital by public health insurance in France. Finally, alive status at a certain time was also reliable because it was searched for between 2 distinct encounters.

Global performances, global performance per hospital, performances per sex and per hospital, and performances per country of birth and per hospital were calculated using the stratified sampling proportion method. To calculate these performances, we needed the percentages of patients born outside of France for the 3 cities. French national census data for 2012 [14] yielded 4.5% (40,394/897,639) for Nantes, 4.3% (29,697/690,618) for Rennes, and 8.4% (97,988/1,166,527) for Lille. For this calculation, we considered half of the population to be composed of men and the other half of women.

Implementation and Execution Time Evaluation

We developed an R package to run on parallel cores and automatically select by default the most efficient number of cores to use depending on the number of records to match, the number of available cores, and the available RAM. The number of cores used still fit in the parameters. We used the packages “future” and “future.apply” to enable Linux and Windows compatibility. We measured the execution time to successively match 200, 2000, 20,000, 200,000, and 2 million patients from the Nantes BDW with 11 million records from the FNDM. We tested various core numbers on 3 cores and 15 GB of RAM (1 core and 1 GB of RAM on the laptop used were left free for the operating system).

Ethical Considerations

Each of the 3 BDWs had a previous authorization from the National Information Science and Liberties Commission. These authorizations included data quality controls that our algorithm contributes to.

Results

Performances of the Matching Algorithms

Table 3 compares the performances between the direct and DLD-based algorithms for all 3 hospitals combined. Sensitivity of the DLD-based algorithm was 11% higher than that of the direct algorithm (93.3%, 95% CI 92.8-93.9 vs 82.7%, 95% CI 81.8-83.6; *P*<.001). Specificity of the DLD-based algorithm was <1% lower than that of the direct algorithm (99%, 95% CI 98.7-99.2 vs 99.9%, 95% CI 99.8-100; *P*<.001). Table 4 presents overall performances by hospital for both algorithms. Sensitivity of the DLD-based algorithm for the Rennes and Lille samples was 12% higher than that for the Nantes sample (85.3%, 95% CI 83.8-86.8 vs 97.3%, 95% CI 96.7-97.9; *P*<.001). Specificity of the DLD-based algorithm was >98% in all samples (98.2% to 99.4%; *P*<.001).

Table 5 presents the performances of the DLD-based algorithm per sex and per hospital. In Lille, sensitivity was equal for both sexes (97.3%; *P*>.99). Sensitivity was higher for men than for women in Nantes (87%, 95% CI 85.1-89.0 vs 83.6%, 95% CI 81.4-85.8; *P*=.006) and in Rennes. In all hospitals, specificity for women (98.6% to 99.6%) was higher than that for men (97.9% to 99.2%), but with no statistically significant differences (*P*>.05).

Table 6 presents performances of the DLD-based algorithm per birth country and per hospital. For every hospital, sensitivity of the DLD-matching algorithm was ~10% higher (*P*<.001) for people born in France than for people born outside France (Nantes: 85.8%, 95% CI 84.3-87.3 vs 74.9%, 95% CI 72.8-77.0; *P*<.001). Specificity was >98% for every sample (range 98.2%-99.8%). In Lille, specificity was equal for people born both in and outside of France (99.4%, 95% CI 99-99.7; *P*<.99). In Nantes and Rennes, specificity for people born out of France (98.8%) was higher than that for people born in France (98.2% to 99.3%; *P*<.05).

Table 3. Global performances of the Damerau-Levenshtein distance (DLD)–based algorithm versus that of the direct-matching algorithm.

Matching algorithm	Sample size, N	Sensitivity, % (95% CI)	Specificity, % (95% CI)
Distance-based[a]	21860	93.3 (92.8-93.9)	99 (98.7-99.2)
Direct	21860	82.7 (81.8–83.6)	99.9 (99.8-100)
P value McNemar test	N/A[b]	<.001	<.001

[a] Maximal total distance: 2.

[b] N/A: not applicable.

Table 4. Global performances of the Damerau-Levenshtein distance (DLD)–based algorithm versus the direct-matching algorithm per university hospital.

University hospital	Total sample, n	Patients born outside of France used for weights, %	Se[a] DLD[b], % (95% CI)	Se direct[c], % (95% CI)	Sp[d] DLD, % (95% CI)	Sp direct, % (95% CI)
Nantes	Se: 3660 Sp: 4000	4.5	85.3 (83.8-86.8)	74.6 (72.8-76.4)	99.3 (99-99.7)	99.9 (99.7-100)
Rennes	Se: 2500 Sp: 4000	4.3	97.3 (96.7-97.9)	86.0 (84.6-87.4)	98.2 (97.7-98.8)	100 (99.9-100)
Lille	Se: 3700 Sp: 4000	8.4	97.3 (96.8-97.9)	87.5 (86.2-88.8)	99.4 (99-99.7)	99.9 (99.8-100)
P value Fisher exact test	N/A[e]	N/A	<.001	<.001	<.001	.01

[a] Se: sensitivity.

[b] DLD: Damerau-Levenshtein distance–based matching algorithm (Maximal total distance used: 2).

[c] Direct: direct-matching algorithm.

[d] Sp: specificity.

[e] N/A: not applicable.

Table 5. Performances of the Damerau-Levenshtein distance (DLD)–based matching algorithm by sex and university hospital.

University hospital	Total sample	Se[a] women, % (95% CI)	Se men, % (95% CI)	*P* value Fisher exact test	Sp[b] women, % (95% CI)	Sp men, % (95% CI)	*P* value Fisher exact test
Nantes	Se women: 1660 Se men: 2000 Sp women: 2000 Sp men: 2000	83.6 (81.4-85.8)	87 (85.1-89)	.006	99.4 (99-99.9)	99.2 (98.7-99.8)	.57
Rennes	Se women: 1300 Se men: 1200 Sp women: 2000 Sp men: 2000	96 (94.9-97.1)	98.6 (98.1-99.2)	<.001	98.6 (97.8-99.3)	97.9 (97.0-98.8)	.12
Lille	Se women: 1700 Se men: 2000 Sp women: 2000 Sp men: 2000	97.3 (96.5-98.1)	97.3 (96.6-98)	>.99	99.6 (99.2-100)	99.1 (98.6-99.6)	.08

[a] Se: sensitivity.

[b] Sp: specificity.

Finally, use of the DLD was more efficient for women and people born outside France than for men and people born in France. In Nantes, an increase from 0 to 2 for the maximal total DLD increased the sensitivity by 1.85% for MBIF, 4.4% for MBOF, 2.9% for WBIF, and 6.6% for WBOF. Performances per sex, birth country, and maximal total DLD for the DLD-based algorithm are available for Nantes hospital in Multimedia Appendix 4.

Details on the performances per strata and repartition of the DLD of the valid (sensitivity) and invalid (specificity) pairs are available for Nantes hospital in Multimedia Appendices 5-6.

Table 6. Performances of the Damerau-Levenshtein distance (DLD)–based matching algorithm per birth country and per university hospital.

University hospital	Sample size	Se[a,b] BIF[c], % (95% CI)	Se[b] BOOF[d], % (95% CI)	*P* value Fisher exact test	Sp[eb] BIF, % (95% CI)	Sp[b] BOOF, % (95% CI)	*P* value Fisher exact test
Nantes	Se, BIF: 2000 Se, BOOF: 1660 Sp, BIF: 2000 Sp, BOOF: 2000	85.8 (84.3-87.3)	74.9 (72.8-77)	<.001	99.3 (98.9-99.7)	99.8 (99.6-100)	.03
Rennes	Se, BIF: 2000 Se, BOOF: 500 Sp, BIF: 2000 Sp, BOOF: 2000	97.8 (97.1-98.4)	87.6 (84.7-90.5)	<.001	98.2 (97.6-98.7)	99.8 (99.5-100)	<.001
Lille	Se, BIF: 2000 Se, BOOF: 1700 Sp, BIF: 2000 Sp, BOOF: 2000	98.3 (97.7-98.9)	86.8 (85.2-88.4)	<.001	99.4 (99-99.7)	99.4 (90-99.7)	>.99

[a] Se: sensitivity.

[b] Max total distance used: 2

[c] BIF: patient born in France.

[d] BOOF: patient born out of France.

[e] Sp: specificity.

Application of the Nantes BDW

Among the 1,974,786 (97%) patients recorded as living in the Nantes BDW, 205,698 (10.4%) were matched to the FNMD. Table 7 presents the sex repartition among these patients, and Table 8 presents the age at death by sex. Among all patients linked to the FNMD, 117,563 (57%) were men, and they died 8 years earlier than women did (age 74 years vs 82 years, respectively).

Table 7. Sex of patients recorded as living in the Nantes biomedical data warehouse (BDW) and linked to the French National Mortality Database (FNMD).

Sex	Patients in Nantes BDW[a] (N=205,698), n (%)
Women	88,090 (42.82)
Men	117,563 (57.15)
Unknown	45 (0.022)

[a] BDW: biomedical data warehouse.

Table 8. Age at death of patients recorded as living in the Nantes biomedical data warehouse (BDW) and linked to the French National Mortality Database (FNMD).

Variable	Women (N=88,090)	Men (N=117,563)	Unknown (N=45)
Death age (years), median (IQR)	82 (20)	74 (21)	69 (21)

Large-scale Testing

On our laptop, the execution time to match 200 patients from BDW with the FNMD was 3 minutes, and it was 78 hours to match 2,000,000 patients from BDW with the FNMD. The execution time per patient decreased with the total number of patients. Details are available in Multimedia Appendix 7. The use of blocking techniques reduced the number of required comparisons by at least 40,000 times.

Open-Access R Code

The R package for the DLD algorithm, called Inseehop, is open access on GitLab [15] and will be maintained and updated by the authors.

Discussion

Background

We developed a large-scale DLD-based record-linkage algorithm to match patients from BDWs in France with the FNMD. We then compared the algorithm's performances with those of a direct-matching algorithm for 3 samples from the Lille, Nantes, and Rennes BDWs.

Performances That Increased Sensitivity/Recall and Reduced Differential Biases

Overall, sensitivity/recall was approximately 11% higher with the DLD-based algorithm than with the direct algorithm. This highlights the importance of advanced data cleaning and knowledge of a naming system through DLD use.

Moreover, sensitivity was approximately 12% higher for the Lille and Rennes evaluation samples than for the Nantes sample, possibly owing to differences in the BDW data quality or less efficient death report management by regional city halls. Hence, when possible, each center interested in reusing our algorithm should compute its own FNMD-matching performance evaluation.

Sensitivity was approximately 3% lower for women than for men in the Nantes and Rennes samples. This may have been because women are more likely to change their surname after marriage, whereas most men do not; thus, women's birth surnames are not always registered. The 2020 Réferentiel d'Identitovigilance National Identity Monitoring Guidelines in France [16] recommendusing patients' birth surnames, even for married women. These differences should consequently disappear in the future.

Sensitivity was higher for people born in France than for those born outside France. This result was expected because other countries' administrations do not send death reports to INSEE when their citizens die on their territory. Another explanation is that the same surname, first name, or middle name of a non-French patient can have multiple translations in French.

These sensitivity differences (per sex, birth country, or hospital) must be considered when performing an analysis to avoid differential biases between groups.

Finally, increasing the maximal total DLD in the DLD-based algorithm reduced performance gaps between men and women and patients born in and outside of France, which helped limit the differential biases between groups. Specificity was ≥98% for both sexes, birth country, and hospital, which greatly reduced the risk of differential biases between groups.

Blocking the birth date and then concatenating the first 4 characters of the first name and the first 4 characters of the surname reduced the time needed to match 2 million patients from ~366 years to 78 hours. However, records from the local database cannot be compared with those from the FNMD if both these blocking criteria differ; they can only be compared if only 1 differs or both are equal.

Death Prevalence Was Greatly Underestimated in the BDWs

Applying the DLD-based algorithm to the Nantes BDW revealed that >200,000 patients registered as alive were actually linked to the FNMD, which was approximately 3 times more than the 60,000 patients initially registered as deceased. More men died, and at younger ages, which is consistent with the actual demographic data discussed earlier in this paper.

Large-scale Matching on a Daily Routine Basis With Minimal Local Computing Capabilities

The program implemented in R software to work on parallel cores was able to run with 2 million patients from the Nantes BDW and 11 million deceased people from the FNMD on 15 GB of RAM and 3 cores in a reasonable duration. Execution time could be improved with higher performance platforms; our laptop was not ideal due to its low computing capabilities and overheating problems during the 3 cores calculations. Because data stayed on hospital computers and no external service was involved, confidentiality was optimal. Moreover, only popular R packages were necessary to run it, which is useful for users who lack administrator rights on their machines.

Quality of the Gold Standard

As described earlier in this paper, our gold standard was reliable both for sensitivity and specificity evaluation because (1) it was completely independent from the FNMD, (2) deceased status in the hospital databases was reliable, and (3) alive status at a certain time was searched between 2 distinct encounters. For some patients in our sample, names, surnames, birth date, or sex may have been incorrect, as in every database. Nevertheless, this was not a problem because our algorithm could manage these kinds of errors.

Additional Use Cases

Although our algorithm was originally conceived for linking a large-scale BDW with the large-scale FNMD, it can be used for other purposes, such as matching a large hospital database with an insurance database.

Limitations

Initially, the expected sample size to evaluate performance at each center was 8000. However, in some cases, there were too few patients with a registered birth country to obtain 1000 patients per strata per center, particularly for WBOF. Nevertheless, sample sizes were sufficient to yield small confidence intervals and significant *P* values.

Another limitation was our methodology, which likely overestimated the sensitivity. Deaths of patients that occurred both inside and outside the hospital and were then communicated to the hospital were not representative of all deceased people. The only way to improve the gold standard would be to conduct an individual investigation of vital status for every patient, which is not possible without significant resources on a large scale.

Conclusions

While matching operations using names are sensitive computational operations, the Inseehop package we released is easy to run on premises, facilitating compliance with local cybersecurity frameworks. The use of advanced deterministic matching algorithm such as the DLD-based algorithm is an insightful example of combining open-source external data to improve the usage value of BDWs.

Acknowledgments

This work was supported in part by the cluster DELPHI - NExT financed by French National Research Agency (Agence Nationale de la Recherche, ANR), integrated to France 2030 plan, by Région Pays de la Loire and by Nantes Métropoles. We thank Drs David Lair, Romain Casey, and Florance Molinié for their suggestions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the data cleaning and algorithms.
[DOC File , 49 KB - medinform_v10i11e36711_app1.doc]

Multimedia Appendix 2
Damerau-Levenshtein distances.
[DOCX File , 23 KB - medinform_v10i11e36711_app2.docx]

Multimedia Appendix 3
Choice of the most pertinent FNMD records for one local database record in three consecutives steps.
[DOCX File , 24 KB - medinform_v10i11e36711_app3.docx]

Multimedia Appendix 4
Performance per sex, birth country and maximal total DLD for the DLD-based algorithm in Nantes.
[DOCX File , 28 KB - medinform_v10i11e36711_app4.docx]

Multimedia Appendix 5
Repartition of the different DLD pair types for a maximal total distance of 5 for the Nantes sample sensitivity estimation.
[DOCX File , 26 KB - medinform_v10i11e36711_app5.docx]

Multimedia Appendix 6
Repartition of the different DLD pair types for a maximal total distance of 5 for the Nantes sample specificity estimation.
[DOCX File , 24 KB - medinform_v10i11e36711_app6.docx]

Multimedia Appendix 7
Total execution time and execution time per 10,000 patients.
[DOCX File , 24 KB - medinform_v10i11e36711_app7.docx]

References

1. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52 [FREE Full text] [doi: 10.15265/IY-2017-007] [Medline: 28480475]
2. Répertoire national d'identification des personnes physiques. Documentation du SNDS. 2019 Oct 23. URL: https://documentation-snds.health-data-hub.fr/glossaire/rnipp.html#contenu [accessed 2022-01-18]
3. Comission d'accès aux documents administratifs. Avis 20182992 - Séance du 17/05/2019. Avis de la comission d'accès aux documents administratifs. 2019 May 17. URL: https://www.cada.fr/20182992 [accessed 2022-01-18]
4. Gu L, Baxter R, Vickers D, Rainsford C. Record Linkage: Current Practice and Future Directions. CSIRO Mathematical and Information Sciences Technical Report 2003 Jun 18:1-32 [FREE Full text]
5. Bounebache K, Quantin C, Benzenine E, Obozinski G, Rey G. Revue Bibliographique des Méthodes de Couplage des Bases de Données : Applications et Perspectives dans le Cas des Données de Santé Publique. Journal de la société française de statistiques 2018 Dec 13;159 n°3:79-123 [FREE Full text]
6. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969 Dec;64(328):1183-1210 [FREE Full text] [doi: 10.1080/01621459.1969.10501049]
7. Copas JB, Hilton FJ. Record linkage: statistical models for matching computer records. J R Stat Soc Ser A Stat Soc 1990;153(3):287-320 [FREE Full text] [Medline: 12159128]
8. Moore CL, Amin J, Gidding HF, Law MG. A new method for assessing how sensitivity and specificity of linkage studies affects estimation. PLoS One 2014 Jul 28;9(7):e103690 [FREE Full text] [doi: 10.1371/journal.pone.0103690] [Medline: 25068293]

9. Fournel I, Schwarzinger M, Binquet C, Benzenine E, Hill C, Quantin C. Contribution of record linkage to vital status determination in cancer patients. Stud Health Technol Inform 2009;150:91-95. [Medline: 19745273]
10. Moussa MD, Lamer A, Labreuche J, Brandt C, Mass G, Louvel P, et al. Mid-term survival and risk factors associated with myocardial injury after fenestrated and/or branched endovascular aortic aneurysm repair. Eur J Vasc Endovasc Surg 2021 Oct;62(4):550-558. [doi: 10.1016/j.ejvs.2021.02.043] [Medline: 33846076]
11. Bannay A, Bories M, Le Corre P, Riou C, Lemordant P, Van Hille P, et al. Leveraging national claims and hospital big data: cohort study on a statin-drug interaction use case. JMIR Med Inform 2021 Dec 13;9(12):e29286 [FREE Full text] [doi: 10.2196/29286] [Medline: 34898457]
12. INSEE. Fichier des personnes décédées. data.gouv.fr. URL: https://www.data.gouv.fr/fr/datasets/fichier-des-personnes-decedees/ [accessed 2022-01-18]
13. BARD GV. Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric. In: Proceedings of the fifth Australasian symposium on ACSW frontiers, volume 68. 2017 Feb 30 Presented at: fifth Australasian symposium on ACSW frontiers; january 2007; Ballarat, Australia p. 117-124.
14. BRUTEL C. La localisation géographique des immigrés. INSEE Première. 2016 Apr 19. URL: https://www.insee.fr/fr/statistiques/2121524 [accessed 2022-01-18]
15. GUARDIOLLE V, BAZOGE A, LAMER A. InseeHop : open source code. Gitlab. 2021 Oct 21. URL: https://gitlab.com/ricdc/insee-deces [accessed 2022-01-18]
16. Direction Générale de l'offre de soins. RÉFÉRENTIEL NATIONAL D'IDENTITOVIGILANCE. 2020. URL: https://tinyurl.com/yc5kfwp5 [accessed 2022-10-06]

Abbreviations

BDW: biomedical data warehouse
DLD: Damerau-Levenshtein distance
INSEE: Institut National de la Statistique et des Études Économiques
FNMD: French National Mortality Database
MBIF: men born in France
MBOF: men born outside France
SNDS: Système National des Données de Santé
WBIF: women born in France
WBOF: women born outside France

Edited by C Lovis; submitted 21.01.22; peer-reviewed by L Hagen; comments to author 27.03.22; revised version received 04.04.22; accepted 11.04.22; published 01.11.22.

Please cite as:
Guardiolle V, Bazoge A, Morin E, Daille B, Toublant D, Bouzillé G, Merel Y, Pierre-Jean M, Filiot A, Cuggia M, Wargny M, Lamer A, Gourraud PA
Linking Biomedical Data Warehouse Records With the National Mortality Database in France: Large-scale Matching Algorithm
JMIR Med Inform 2022;10(11):e36711
URL: https://medinform.jmir.org/2022/11/e36711
doi:10.2196/36711
PMID:36318244

Original Paper

Discovery and Analytical Validation of a Vocal Biomarker to Monitor Anosmia and Ageusia in Patients With COVID-19: Cross-sectional Study

Eduardo Higa[1*], BSc, MSc; Abir Elbéji[1*], BSc, MSc; Lu Zhang[2*], BSc, MSc; Aurélie Fischer[1*], BSc, MSc; Gloria A Aguayo[1*], MD, PhD; Petr V Nazarov[2*], BSc, MSc, PhD; Guy Fagherazzi[1*], BSc, MSc, DPhil

[1]Deep Digital Phenotyping Research Unit, Department of Population Health, Luxembourg Institute of Health, Strassen, Luxembourg

[2]Bioinformatics Platform, Quantitative Biology Unit, Luxembourg Institute of Health, Strassen, Luxembourg

[*]all authors contributed equally

Corresponding Author:
Guy Fagherazzi, BSc, MSc, DPhil
Deep Digital Phenotyping Research Unit
Department of Population Health
Luxembourg Institute of Health
1A-B, rue Thomas Edison
Strassen, L1445
Luxembourg
Phone: 1 26970 457
Fax: 1 26970 719
Email: guy.fagherazzi@gmail.com

Abstract

Background: The COVID-19 disease has multiple symptoms, with anosmia and ageusia being the most prevalent, varying from 75% to 95% and from 50% to 80% of infected patients, respectively. An automatic assessment tool for these symptoms will help monitor the disease in a fast and noninvasive manner.

Objective: We hypothesized that people with COVID-19 experiencing anosmia and ageusia had different voice features than those without such symptoms. Our objective was to develop an artificial intelligence pipeline to identify and internally validate a vocal biomarker of these symptoms for remotely monitoring them.

Methods: This study used population-based data. Participants were assessed daily through a web-based questionnaire and asked to register 2 different types of voice recordings. They were adults (aged >18 years) who were confirmed by a polymerase chain reaction test to be positive for COVID-19 in Luxembourg and met the inclusion criteria. Statistical methods such as recursive feature elimination for dimensionality reduction, multiple statistical learning methods, and hypothesis tests were used throughout this study. The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Prediction Model Development checklist was used to structure the research.

Results: This study included 259 participants. Younger (aged <35 years) and female participants showed higher rates of ageusia and anosmia. Participants were aged 41 (SD 13) years on average, and the data set was balanced for sex (female: 134/259, 51.7%; male: 125/259, 48.3%). The analyzed symptom was present in 94 (36.3%) out of 259 participants and in 450 (27.5%) out of 1636 audio recordings. In all, 2 machine learning models were built, one for Android and one for iOS devices, and both had high accuracy—88% for Android and 85% for iOS. The final biomarker was then calculated using these models and internally validated.

Conclusions: This study demonstrates that people with COVID-19 who have anosmia and ageusia have different voice features from those without these symptoms. Upon further validation, these vocal biomarkers could be nested in digital devices to improve symptom assessment in clinical practice and enhance the telemonitoring of COVID-19–related symptoms.

Trial Registration: Clinicaltrials.gov NCT04380987; https://clinicaltrials.gov/ct2/show/NCT04380987

(JMIR Med Inform 2022;10(11):e35622) doi:10.2196/35622

KEYWORDS

vocal biomarker; COVID-19; ageusia; anosmia; loss of smell; loss of taste; digital assessment tool; digital health; medical informatics; telehealth; telemonitoring; biomarker; pandemic; symptoms; tool; disease; noninvasive; AI; artificial intelligence; digital; device

Introduction

In the context of the COVID-19 pandemic, declared by the World Health Organization in early March 2020, the fast and easy diagnosis of the disease has become an important concern. Anosmia, an olfactory dysfunction that leads to a temporary or permanent loss of olfaction, is present in 75% to 95% [1-3] of infected patients, whereas ageusia, a gustatory dysfunction resulting from the loss of functions of the tongue, is present in 50% to 80% [1,2,4,5] of infected people and can predict infection [6], depending on the virus strain and population characteristics. Proportionally, younger and female patients showed higher rates of these symptoms—a proven correlation due to differences in cytokine storms [5,7].

Monitoring these symptoms is highly needed and could be facilitated with an easy-to-use digital health solution. In individual who are infected but not tested, checking such symptoms could also serve as a rapid screening solution and suggest the realization of a test to limit the spread of the virus. There are also many concerns about the so-called Long COVID, where anosmia and ageusia are frequently reported [8]. A fast, noninvasive symptom assessment tool would be useful to better understand the whole spectrum of the disease and monitor Long COVID's evolution over time. Furthermore, these symptoms are associated with neurodegenerative diseases such as Parkinson and Alzheimer diseases [9,10] and can lead to multiple impacts, such as nutritional deficits [11].

The human voice is a rich medium that serves as a primary source of communication between individuals. Furthermore, talking is a uniquely human ability; it is one of the most natural and energy-efficient ways of interacting with each other. Slight alterations, for instance, due to a COVID-19–related symptom, are made by changes either in respiration, phonation, or articulation—the 3-stage process of voice production [12]—which will result in variations of pitch, tone, fundamental frequency, and many other aspects of our voice. Recent developments in audio signal processing and artificial intelligence methods have enabled a more refined and in-depth voice features analysis that surpasses the human level of perception and can solve complex problems in the health care domain.

This study aimed to test the hypothesis that anosmia and ageusia following a SARS-CoV-2 infection can result in modifications in voice production that could help detect and monitor these specific symptoms. To achieve our objective, we used data from the prospective Predi-COVID cohort study, where both voice and COVID-19–related symptoms were frequently recorded. We analyzed voice signals, built panels of vocal biomarkers, and internally validated them using the developed prediction models.

Methods

Study Population

This study used data from the Predi-COVID cohort [13]—a prospective, hybrid cohort started in May 2020 composed of adult patients (aged >18 years) who were confirmed, by a polymerase chain reaction test, to be positive for COVID-19 in Luxemburg, both in and out of the hospital.

The first contact with potential participants was made via phone by collaborators from the Health Inspection. Those who agreed to take part were contacted by an experienced nurse or clinical research associate from the Clinical and Epidemiological Investigation Center, who explained the study and organized visits at home or the hospital, and informed consent for participation was obtained.

Through the first 14 days following inclusion, participants were assessed daily through a web-based questionnaire. A subcohort agreed to be digitally followed by a digital app that was dedicated to voice recording in cohort studies. To guarantee a minimum quality standard, participants were instructed to register the audio in a calm place while keeping a specific distance from the microphone**.** An audio example of what was expected was also available.

Each day, 2 types of voice recordings were performed. In the first recording, called Type 1 audio, participants had to read an extract from the Declaration of Human Rights, Article 25, paragraph 1 (Multimedia Appendix 1) in their preferred language: French, German, English, or Portuguese; and in the second recording, called Type 2 audio, they were asked to hold the “[a]” vowel phonation without breathing as long as they could. For this analysis, we considered only voice recordings from the first 2 weeks after inclusion where the symptoms were collected regularly. Since the study is in a real-life setting, the number of vocal samples per participant may have differed.

Ethics Approval

The study was approved by the National Research Ethics Committee of Luxembourg (study 202003/07) in April 2020 and is registered on ClinicalTrials.gov (NCT04380987).

Inclusion Criteria

All participants who had no missing data on sex, information on the studied outcome, and both types of audio recordings on the same day during the first 14 days of follow-up were included in the model.

Anosmia and Ageusia

In this study, both anosmia and ageusia were the outcomes and were united in a single variable based on the participant’s perception. The specific question was the following: “Did you notice a strong decrease or a loss of taste or smell?” The possible answers were “yes” or “no.” Since the loss of smell can substantially affect taste functions [14], uniting the 2 symptoms is expected to be a more realistic strategy because the outcome is self-reported, and it would not be easy for the participant to clearly distinguish between ageusia and anosmia.

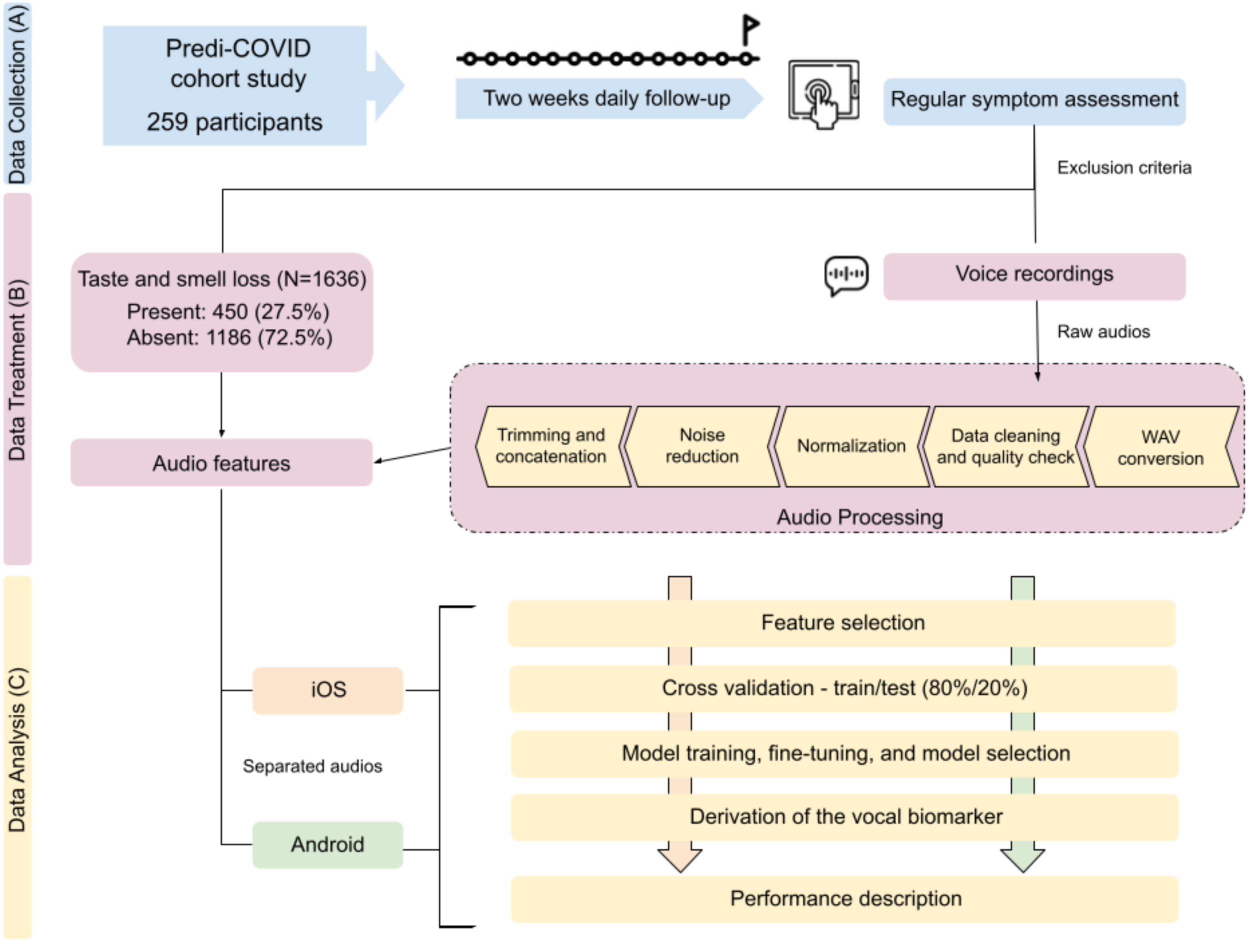
Prediction Data

The prediction models were based on both Type 1 and Type 2 voice recordings to predict the outcome. To maximize the information given to the model, both types were concatenated and used as a single input to the learning model. The audio format and recording settings varied depending on the operating system of the smartphone used to record it: Android devices were registered in 3gp format, whereas iOS devices were registered in m4a format. These 2 formats were also analyzed separately to create predictive models for each type of operating system.

Voice Signal Treatment

The audios were preprocessed to remove poorly recorded or corrupted files, and the remaining ones were then normalized and cleaned for noise. Type 1 and Type 2 audios were both sampled with an 8000 Hz sample rate, as different rates did not significantly improve the model. Audios were then concatenated, which resulted in a final sample from which the features were extracted. The pipeline can be found in Figure 1.

Figure 1. Learning pipeline to the discovery of biomarkers. (A) Data collection from Predi-COVID and exclusion criteria. (B) Data treatment of audio data and studied outcome. (C) Data analysis for both audio formats done in parallel.



OpenSMILE

The Munich Open-Source Media Interpretation by Large Feature-Space Extraction (openSMILE) is a modular and flexible research-only toolkit for extracting features for signal processing and machine learning applications. It is widely used in the speech recognition community, the area of affective computing, and music information retrieval [15]. The package provides many functionalities, such as windowing functions, resampling, and fast Fourier transform. It can extract a wide range of features including frame energy, Mel-frequency cepstral coefficients, loudness, jitter, shimmer, and many others. The specific openSMILE feature set is the same as that used in The Interspeech 2016 Computational Paralinguistics Challenge [16], originally chosen to assess sentiments through the voice. Within it, there are 2 feature levels: functionals, which gather much more detailed information and reach up to 6473 different features; and low-level descriptors, measures that are closely related to the signal and reach up to 66 features [17]. The latter feature level is embedded in the functional features, and the full set of feature categories is shown in Multimedia Appendix 2.

Recursive Feature Elimination

Recursive feature elimination (RFE) is a dimensionality reduction method that recursively ranks features according to a measure of importance defined by another classifier (linear regression and random forests, for example), and at each iteration, the ones with the lowest rank are removed until the desired number is reached [18]. The minimum number of features was set to 10, a linear regression was used to define the weights, and 25 features were removed at each iteration (step=25). This process was performed using 10-fold cross-validation.

Statistical Analysis Methods

Chi-square test and Student *t* test (2-tailed) were used in this study. We applied standard machine learning algorithms that work with structured data to analyze the extracted features. Random forests [19], k-nearest neighbors (KNN) [20], and support vector machines [21] were used to avoid biases from a single predictor and test different approaches on the same data.

All hyperparameters were hyper tuned using grid search from *scikit-learn* (version 0.22.2) [22], maximizing the weighted area under the receiver operating characteristic curve (ROC AUC). The data were divided into a 60%/20%/20% proportion for training, validation, and testing, respectively. To evaluate its sensibility, 10-fold cross-validation was first performed on the training set to analyze the dispersion of the metrics, and then the final model was built on the testing set.

The final model was chosen based on the following metrics: precision, recall, *F*-measure, and accuracy. Given the nature of the problem, we assumed that having false negatives was worse than having false positives, since one can develop severe symptoms and continue to spread the virus if misclassified, so the recall for those positive to the studied outcome should be maximized. The weighted ROC AUC was also taken into account since it indicates the overall performance of the model in terms of its accuracy at various diagnostic thresholds used to discriminate between 2 classes [23].

To derive the vocal biomarker from the prediction model, we used the final probability of being classified as having anosmia or ageusia; its distribution was further evaluated in both groups.

Results

Descriptive Data

After excluding all data that did not meet the inclusion criteria, we used descriptive statistics to characterize the study participants. The final study population had a total of 259 participants, and age, sex, and BMI were associated with the outcome (*P*<.001, *P*<.001, *P*<.001, respectively). Younger (aged <35 years) and female participants showed higher rates of ageusia and anosmia.

Participants were aged 41 (SD 13) years on average with a BMI of 25.4 (SD 4.6)—the intersection between normal weight and overweight [24]. Antibiotics intake, asthma, and smoking were highly unbalanced clinical features (present in n=29, 11.2%; n=10, 3.9%; and n=177, 68.3% of participants, respectively). The data set was balanced for sex (female: n=134, 51.7%; male: n=125, 48.3%), and the analyzed symptom was present in 94 (36.3%) out of 259 participants and in 450 (27.5%) out of 1636 of audio recordings. This result occurs due to a variation in the number of recordings per participant, with each one having an average of 6 audio recordings. Finally, Type 1 audio had an average length of 28.5 s, whereas Type 2 audio had an average length of 18.9 s.

As the audio format was linearly separable when analyzing the outcome, shown in Figure 2, they were separated in the analysis. When divided by audio format, no significant difference was found between the 2 sets of participants. Clinical features and audio data can be seen in Tables 1-2.

Figure 2. Sample plot with linear separation between 3gp and m4a audio formats. Principal component analysis was used on the extracted features, and the first 2 dimensions were used to plot the samples.

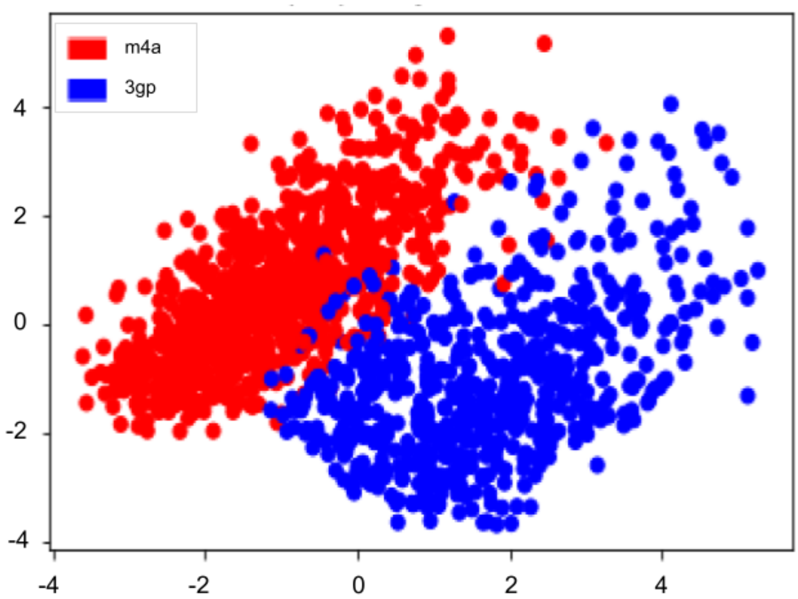


Table 1. Description of the participants, containing clinical data to characterize the general population of the study and the loss of smell and taste. All categorical data are represented as the total number and its percentage.

Description	Total (N=259)	Audio format, operating system		*P* value[a]
		m4a, iOS (n=161)	3gp, Android (n=98)	
Symptom, n (%)				.51
Normal taste and smell	165 (63.7)	105 (65.2)	60 (61.2)	
Loss of taste and smell	94 (36.3)	56 (34.8)	38 (38.8)	
Sex, n (%)				.14
Female	134 (51.7)	89 (55.3)	45 (45.9)	
Male	125 (48.3)	72 (44.7)	53 (54.1)	
Antibiotic, n (%)				.42
No	230 (88.8)	141 (87.6)	89 (90.8)	
Yes	29 (11.2)	20 (12.4)	9 (9.2)	
Asthma, n (%)				.88
No	249 (96.1)	155 (96.3)	94 (95.9)	
Yes	10 (3.9)	6 (3.7)	4 (4.1)	
Smoking, n (%)				.85
Yes	177 (68.3)	112 (69.6)	65 (66.3)	
Never	44 (17)	26 (16.1)	18 (18.4)	
Former smoker	38 (14.7)	23 (14.3)	15 (15.3)	
Age (years), mean (SD)	40.6 (12.7)	40.6 (13.4)	40.7 (11.5)	.93
BMI (kg/m²), mean (SD)	25.4 (4.6)	25.4 (4.9)	25.5 (4.1)	.80

[a] All *P* values were calculated through chi-square or Student *t* test between m4a and 3gp formats.

Table 2. Description of the audio samples, with their general information.

Description	Total (N=1636)	Audio format, operating system		*P* value[a]
		m4a, iOS (n=999)	3gp, Android (n=637)	
Audio samples per symptom, n (%)				.06
Normal taste and smell	1186 (72.5)	741 (74.2)	445 (69.9)	
Loss of taste and smell	450 (27.5)	258 (25.8)	192 (30.1)	
Number of audio samples per participant, mean (SD)	6.3 (4.5)	6.2 (4.4)	6.5 (4.6)	—[b]
Text reading duration (s), mean (SD)	28.5 (4.1)	28.3 (4.1)	28.9 (4.2)	—
Vowel phonation duration (s), mean (SD)	18.9 (6.8)	18.2 (6.6)	20 (7.1)	—

[a] All *P* values were calculated through chi-square or Student *t* test between m4a and 3gp formats.

[b] Not available.

Feature Extraction

We extracted 6473 features from the concatenated audios. Constant features throughout all the audios were removed from the analysis (50 for Android and 49 for iOS). A RFE method was used to find the best number of features (Multimedia Appendix 3). For 3gp and m4a audios, we selected 3248 and 849 features, respectively.

After extraction, a density plot for the low-level descriptors was made, as shown in Multimedia Appendices 4-5. It can be seen that the distribution of the variables varies depending on the outcome, which reinforces the hypothesis that there are vocal changes related to COVID-19 infection.

Prediction Models’ Performances

The algorithms were first hyper tuned and then trained on all the extracted features and the ones selected through *RFECV*. All models used an 80%/20% stratified proportion for training and testing, respectively, and 10-fold cross-validation was used to assess its sensitivity. The *numpy* seed and the random state of all processes were set to 42 to assure reproducibility, and the samples were weighted to correct the models for unbalanced data.

Models trained on all features had an overall lower performance than those trained with selected features, mainly due to the removal of noise and correlated features (complementary information). The final models for the 3 tested learning algorithms are shown in Table 3. For both formats of audio, we identified KNN as the best method—showing better performances. The AUC was used to choose the best algorithm, and in the end, 3gp had an AUC of 87%, whereas m4a had an AUC of 80%. The specific hyperparameters for each algorithm can be found in Multimedia Appendix 6.

The final models for classifying the loss of taste and smell were KNN for both audio formats and presented a good weighted precision (88% for Android and 85% for iOS), weighted recall (88% for Android and 85% for iOS), and weighted AUC (87% for Android and 80% for iOS). The main difference between the 2 final models is on the recall for the symptomatic class, which was to be maximized (82% for Android and 69% for iOS).

The final vocal biomarker of loss of taste consisted of the probability of being classified as having the symptoms, calculated from the combination of all features selected for each audio format. Its range is shown in Figure 3A, and there was a significant difference between the distribution of probabilities for both 3gp and m4a formats (*P*<.001 and *P*<.001 respectively), which confirms that the model can statistically distinguish the 2 possible conditions, as the probability distribution differs between outcomes.

Figure 3 also presents the confusion matrix for the best classifiers, which shows that they are slightly better in correctly classifying the absence of symptoms than its presence. Additionally, the ROC AUC for each best model is plotted, proving its good learning thresholds.

Table 3. Performance for the 3 different learning methods for each audio format[a].

Audio format (number of selected features), algorithm	Weighted precision	Weighted recall	Recall 1	Accuracy	Weighted AUC[b]	10-fold AUC (SD)
3gp (n=3248)						
KNN[c]	*0.88*	*0.88*	*0.82*	*0.88*	*0.87*	*0.89 (0.05)*
Random forest	0.77	0.77	0.33	0.77	0.64	0.86 (0.03)
SVM[d]	0.81	0.81	0.64	0.81	0.76	0.87 (0.03)
m4a (n=849)						
KNN	*0.85*	*0.85*	*0.69*	*0.85*	*0.80*	*0.89 (0.01)*
Random Forest	0.75	0.77	0.30	0.78	0.70	0.76 (0.02)
SVM	0.78	0.79	0.52	0.79	0.70	0.90 (0.01)

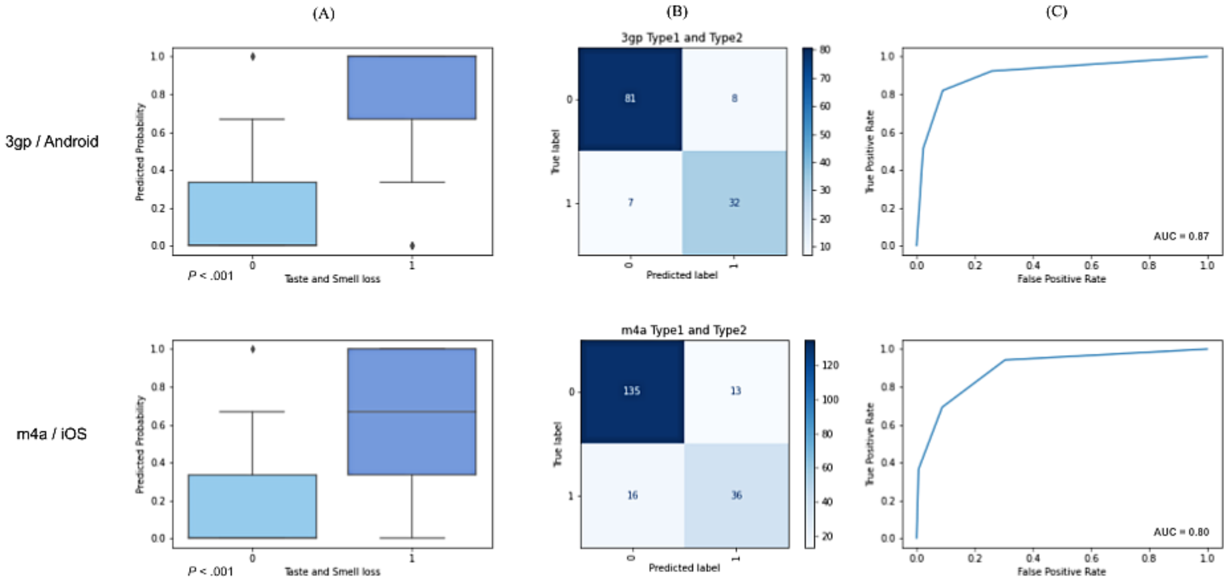
[a] The final model was selected using weighted AUC and is highlighted in italics. Cross-validation was used in the training set as a validation method, and the final model on the testing set showed good adherence to it. The other differences in k-fold and weighted AUC are due to differences in the testing and training set sizes.

[b] AUC: area under the curve.

[c] KNN: k-nearest neighbors.

[d] SVM: support vector machines.

Figure 3. Final models for each audio format. (A) Biomarkers and *P* values from two-sided student's t-test for the presence of anosmia and ageusia were calculated using the probability of classifying as positive. (B) Confusion matrix of the best model. (c) ROC AUC curve. Class 0 represents absence of symptoms and Class 1 the presence of it. ROC AUC: area under the receiver operating characteristic curve.



Discussion

Principal Findings

In this study, we trained artificial intelligence–based algorithms to predict the presence of ageusia and anosmia in patients with COVID-19. In total, 2 predictive models were created based on each smartphone operating system (iOS or Android). We derived 2 sets of vocal biomarkers from these predictive models that should be used together as a single classifier. The biomarkers were then calculated and, after an external validation, can be used to accurately identify patients who present a loss of taste and smell.

Biological Background

Voice is a proven source of medical information, can be easily recorded on a large scale through smart devices [25], and can be easily used to build personalized corpora [26]. Studies have shown great results in the early diagnosis of neurological disorders such as Parkinson disease [27,28], Alzheimer disease [29], and mild cognitive impairment [30,31], since they directly alter the voice, but also in nonneurological conditions such as cardiometabolic [32] and pulmonary [33] diseases. It is important to note that the analysis in this study is new since examples in the literature only analyze short audios (shorter than 5 s) and usually use coughs and other sources of sound [34-36].

Anosmia and ageusia are common COVID-19 symptoms that usually emerge after 5 days of infection [37]. The upper part of the respiratory tract, mainly the olfactory epithelium, is rich in ACE2 and TMPRSS2, 2 main SARS-CoV-2 receptors [38]. Olfactory sensory neurons, on the other hand, were not found to express these receptors, which indicates that the disease itself probably does not directly alter the mechanisms of smell and taste. The infection of support cells, mainly sustentacular and Bowman glands, of these regions and their subsequent malfunction result in alterations in the environment, causing local neuronal death and the final symptom of loss of taste and smell [38,39].

Given that there is no neuronal causality between the loss of taste and smell and voice production, the main pathway in the voice likely involves mechanical influences of COVID-19 infection. The disease alters various systems, such as the respiratory, cardiovascular, and gastrointestinal systems, that if impaired, can directly impact voice characteristics. In mild cases, general symptoms frequently associated with the loss of taste and smell such as dry coughs, insufficient airflow, and pulmonary status also directly affect the production of sounds, resulting in variations that can be used to predict the loss of taste and smell [12].

Strengths and Limitations

The main strengths of this study come from the fact that all participants were confirmed to be positive for COVID-19 by a polymerase chain reaction test. Besides, the majority of the published studies relied on data from hospitalized patients. Therefore, having a cohort of participants mostly at home brings complementary information on the entire spectrum of the disease severity of COVID-19 (from asymptomatic to severe cases). The audio recording is based on a standardized text that has an official translation in many languages, which ensures the high reproducibility of the task in future studies in other countries. The second audio type is a sustained vowel and is, therefore, language-independent and allows analysis without risks of biases due to different articulatory factors, speaking rates, stress, intonations, or any other characteristics that may vary between languages.

This study also has limitations. The recordings are performed in a real-life, noncontrolled environment, which may increase the variability in the quality of the voice recordings. However, since the ultimate objective is to deploy a digital health solution, we cannot rely on well-controlled audio recordings based on a unique device to train the algorithms and should integrate from scratch the diversity of devices and audio recording environments. This study integrates a mixture of different languages in the cohort, but the developed vocal biomarkers cannot be applied to other languages yet. Even though the text is the same, different languages and accents might result in different model performances. Additional external validation studies in other populations that are not well represented in this study (young people) are required at this stage.

In conclusion, we demonstrated that people with COVID-19 who had anosmia and ageusia had different voice features and that it is feasible to accurately predict the presence or absence of this frequent COVID-19 symptom with just a few seconds of the individual's voice. The derived vocal biomarker is strongly associated with the presence of the symptom and could soon be integrated into digital health solutions to help clinicians enhance their consultations or in telemonitoring solutions for remote monitoring. Further external validation studies in other populations and languages are now required.

Acknowledgments

We thank all participants that agreed to be involved in the study, members that collaborated on the launch and monitoring of the Predi-COVID cohort, as well as its scientific committee, the IT team responsible for the development of the app, and the nurses in charge of recruitment, data collection, and management on the field.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Standardized, prespecified text to be read by study participants to collect voice recordings.
[PDF File (Adobe PDF File), 49 KB - medinform_v10i11e35622_app1.pdf]

Multimedia Appendix 2
OpenSMILE categories of extracted features for the 2 feature levels. A detailed description can be found on the web [15].
[PDF File (Adobe PDF File), 30 KB - medinform_v10i11e35622_app2.pdf]

Multimedia Appendix 3
Variation of the AUC performance when varying the number of selected features using RFECV. AUC: area under the curve.
[PNG File , 52 KB - medinform_v10i11e35622_app3.png]

Multimedia Appendix 4
Density plot of the low-level descriptors for 3gp audio format.
[PNG File , 1022 KB - medinform_v10i11e35622_app4.png]

Multimedia Appendix 5
Density plot of the low-level descriptors for m4a audio format.
[PNG File , 873 KB - medinform_v10i11e35622_app5.png]

Multimedia Appendix 6
Hyperparameters for the best algorithms. The random state seed was always set to 42 and the maximum number of iterations to 10000. The implementation of scikit-learn (version 0.22.2) was used.
[PDF File (Adobe PDF File), 29 KB - medinform_v10i11e35622_app6.pdf]

References

1. Samaranayake LP, Fakhruddin KS, Panduwawala C. Sudden onset, acute loss of taste and smell in coronavirus disease 2019 (COVID-19): a systematic review. Acta Odontol Scand 2020 Aug;78(6):467-473. [doi: 10.1080/00016357.2020.1787505] [Medline: 32762282]
2. Ibekwe TS, Fasunla AJ, Orimadegun AE. Systematic review and meta-analysis of smell and taste disorders in COVID-19. OTO Open 2020 Sep 11;4(3):2473974X20957975 [FREE Full text] [doi: 10.1177/2473974X20957975] [Medline: 32964177]
3. Passali GC, Bentivoglio AR. Comment to the article "olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study". Eur Arch Otorhinolaryngol 2020 Aug;277(8):2391-2392 [FREE Full text] [doi: 10.1007/s00405-020-06024-5] [Medline: 32383095]

4. Huang N, Pérez P, Kato T, Mikami Y, Okuda K, Gilmore RC, NIH COVID-19 Autopsy Consortium, HCA OralCraniofacial Biological Network, et al. SARS-CoV-2 infection of the oral cavity and saliva. Nat Med 2021 May;27(5):892-903 [FREE Full text] [doi: 10.1038/s41591-021-01296-8] [Medline: 33767405]
5. Lechien JR, Chiesa-Estomba CM, De Siati DR, Horoi M, Le Bon SD, Rodriguez A, et al. Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): a multicenter European study. Eur Arch Otorhinolaryngol 2020 Aug;277(8):2251-2261 [FREE Full text] [doi: 10.1007/s00405-020-05965-1] [Medline: 32253535]
6. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 2020 Jul;26(7):1037-1040 [FREE Full text] [doi: 10.1038/s41591-020-0916-2] [Medline: 32393804]
7. Lefèvre N, Corazza F, Valsamis J, Delbaere A, De Maertelaer V, Duchateau J, et al. The number of X chromosomes influences inflammatory cytokine production following toll-like receptor stimulation. Front Immunol 2019 May 09;10:1052 [FREE Full text] [doi: 10.3389/fimmu.2019.01052] [Medline: 31143188]
8. Crook H, Raza S, Nowell J, Young M, Edison P. Long covid-mechanisms, risk factors, and management. BMJ 2021 Jul 26;374:n1648. [doi: 10.1136/bmj.n1648] [Medline: 34312178]
9. Rebholz H, Braun RJ, Ladage D, Knoll W, Kleber C, Hassel AW. Loss of olfactory function—early Indicator for COVID-19, other viral infections and neurodegenerative disorders. Front Neurol 2020 Oct 26;11:569333 [FREE Full text] [doi: 10.3389/fneur.2020.569333] [Medline: 33193009]
10. Kovács T. Mechanisms of olfactory dysfunction in aging and neurodegenerative disorders. Ageing Res Rev 2004 Apr;3(2):215-232. [doi: 10.1016/j.arr.2003.10.003] [Medline: 15177056]
11. Kershaw JC, Mattes RD. Nutrition and taste and smell dysfunction. World J Otorhinolaryngol Head Neck Surg 2018 Mar;4(1):3-10 [FREE Full text] [doi: 10.1016/j.wjorl.2018.02.006] [Medline: 30035256]
12. Asiaee M, Vahedian-Azimi A, Atashi SS, Keramatfar A, Nourbakhsh M. Voice quality evaluation in patients with COVID-19: an acoustic analysis. J Voice 2020 Oct 01 [FREE Full text] [doi: 10.1016/j.jvoice.2020.09.024] [Medline: 33051108]
13. Fagherazzi G, Fischer A, Betsou F, Vaillant M, Ernens I, Masi S, et al. Protocol for a prospective, longitudinal cohort of people with COVID-19 and their household members to study factors associated with disease severity: the Predi-COVID study. BMJ Open 2020 Nov 23;10(11):e041834 [FREE Full text] [doi: 10.1136/bmjopen-2020-041834] [Medline: 33234656]
14. Tanasa IA, Manciuc C, Carauleanu A, Navolan DB, Bohiltea RE, Nemescu D. Anosmia and ageusia associated with coronavirus infection (COVID-19) - what is known? Exp Ther Med 2020 Sep;20(3):2344-2347 [FREE Full text] [doi: 10.3892/etm.2020.8808] [Medline: 32765712]
15. openSMILE 3.0. audEERING. URL: https://www.audeering.com/opensmile/ [accessed 2021-07-01]
16. Schuller B, Steidl S, Batliner A, Hirschberg J, Burgoon JK, Baird A, et al. The INTERSPEECH 2016 Computational Paralinguistics Challenge: deception, sincerity & native language. 2016 Presented at: Interspeech 2016; September 8-12, 2016; San Francisco, CA p. 2001-2005. [doi: 10.21437/interspeech.2016-129]
17. openSMILE Python. audEERING. URL: https://audeering.github.io/opensmile-python/ [accessed 2021-07-01]
18. Tang Y, Zhang YQ, Huang Z. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Trans Comput Biol Bioinform 2007 Aug 13;4(3):365-381. [doi: 10.1109/TCBB.2007.70224] [Medline: 17666757]
19. sklearn.ensemble.RandomForestClassifier. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [accessed 2021-08-01]
20. 1.6. Nearest neighbors. scikit-learn. URL: https://scikit-learn.org/stable/modules/neighbors.html [accessed 2021-08-01]
21. 1.4. Support vector machines. scikit-learn. URL: https://scikit-learn.org/stable/modules/svm.html [accessed 2021-08-01]
22. Machine learning in Python. scikit-learn. URL: https://scikit-learn.org/stable/ [accessed 2021-08-01]
23. Walter SD. The partial area under the summary ROC curve. Stat Med 2005 Jul 15;24(13):2025-2040. [doi: 10.1002/sim.2103] [Medline: 15900606]
24. Body mass index (BMI). World Health Organization. URL: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index [accessed 2021-06-01]
25. VynZ Research. Global voice assistant market is set to reach USD 5,843.8 million by 2024, observing a CAGR of 27.7% during 2019–2024: VynZ Research. Globe Newswire. 2020 Jan 28. URL: https://tinyurl.com/5n98af6h [accessed 2021-08-01]
26. Diaz-Asper C, Chandler C, Turner RS, Reynolds B, Elvevåg B. Acceptability of collecting speech samples from the elderly via the telephone. Digit Health 2021 Apr 17;7:20552076211002103 [FREE Full text] [doi: 10.1177/20552076211002103] [Medline: 33953936]
27. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform 2020 Apr;104:103362 [FREE Full text] [doi: 10.1016/j.jbi.2019.103362] [Medline: 31866434]
28. Arora S, Visanji NP, Mestre TA, Tsanas A, AlDakheel A, Connolly BS, et al. Investigating voice as a biomarker for leucine-rich repeat kinase 2-associated Parkinson's disease. J Parkinsons Dis 2018 Oct 17;8(4):503-510. [doi: 10.3233/JPD-181389] [Medline: 30248062]

29. Ahmed S, Haigh AF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 2013 Dec;136(Pt 12):3727-3737 [FREE Full text] [doi: 10.1093/brain/awt269] [Medline: 24142144]
30. Toth L, Hoffmann I, Gosztolya G, Vincze V, Szatloczki G, Banreti Z, et al. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr Alzheimer Res 2018;15(2):130-138 [FREE Full text] [doi: 10.2174/1567205014666171121114930] [Medline: 29165085]
31. Martínez-Sánchez F, Meilán JJG, Carro J, Ivanova O. A Prototype for the voice analysis diagnosis of Alzheimer's disease. J Alzheimers Dis 2018 Jun 19;64(2):473-481. [doi: 10.3233/JAD-180037] [Medline: 29914025]
32. Maor E, Sara JD, Orbelo DM, Lerman LO, Levanon Y, Lerman A. Voice signal characteristics are independently associated with coronary artery disease. Mayo Clin Proc 2018 Jul;93(7):840-847. [doi: 10.1016/j.mayocp.2017.12.025] [Medline: 29656789]
33. Sara JDS, Maor E, Borlaug B, Lewis BR, Orbelo D, Lerman LO, et al. Non-invasive vocal biomarker is associated with pulmonary hypertension. PLoS One 2020 Apr 16;15(4):e0231441 [FREE Full text] [doi: 10.1371/journal.pone.0231441] [Medline: 32298301]
34. Ni X, Ouyang W, Jeong H, Kim J, Tzaveils A, Mirzazadeh A, et al. Automated, multiparametric monitoring of respiratory biomarkers and vital signs in clinical and home settings for COVID-19 patients. Proc Natl Acad Sci U S A 2021 May 11;118(19):e2026610118 [FREE Full text] [doi: 10.1073/pnas.2026610118] [Medline: 33893178]
35. Shimon C, Shafat G, Dangoor I, Ben-Shitrit A. Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires. J Acoust Soc Am 2021 Feb;149(2):1120 [FREE Full text] [doi: 10.1121/10.0003434] [Medline: 33639822]
36. Jayalakshmy S, Sudha GF. Scalogram based prediction model for respiratory disorders using optimized convolutional neural networks. Artif Intell Med 2020 Mar;103:101809. [doi: 10.1016/j.artmed.2020.101809] [Medline: 32143805]
37. Santos REA, da Silva MG, do Monte Silva MCB, Barbosa DAM, Gomes ALDV, Galindo LCM, et al. Onset and duration of symptoms of loss of smell/taste in patients with COVID-19: a systematic review. Am J Otolaryngol 2021 Mar;42(2):102889 [FREE Full text] [doi: 10.1016/j.amjoto.2020.102889] [Medline: 33445036]
38. Brann DH, Tsukahara T, Weinreb C, Lipovsek M, Van den Berge K, Gong B, et al. Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. Sci Adv 2020 Jul 31;6(31):eabc5801 [FREE Full text] [doi: 10.1126/sciadv.abc5801] [Medline: 32937591]
39. Meunier N, Briand L, Jacquin-Piques A, Brondel L, Pénicaud L. COVID 19-induced smell and taste impairments: putative impact on physiology. Front Physiol 2020 Jan 26;11:625110 [FREE Full text] [doi: 10.3389/fphys.2020.625110] [Medline: 33574768]

Abbreviations

KNN: k-nearest neighbors
openSMILE: Open-Source Media Interpretation by Large Feature-Space Extraction
RFE: recursive feature elimination
ROC AUC: area under the receiver operating characteristic curve

Edited by C Lovis; submitted 10.12.21; peer-reviewed by K Rahmani; comments to author 26.07.22; revised version received 11.08.22; accepted 07.09.22; published 08.11.22.

Please cite as:
Higa E, Elbéji A, Zhang L, Fischer A, Aguayo GA, Nazarov PV, Fagherazzi G
Discovery and Analytical Validation of a Vocal Biomarker to Monitor Anosmia and Ageusia in Patients With COVID-19: Cross-sectional Study
JMIR Med Inform 2022;10(11):e35622
URL: https://medinform.jmir.org/2022/11/e35622
doi:10.2196/35622
PMID:36265042

Original Paper

Motion Artifact Reduction in Electrocardiogram Signals Through a Redundant Denoising Independent Component Analysis Method for Wearable Health Care Monitoring Systems: Algorithm Development and Validation

Fabian Andres Castaño Usuga[1], PhD; Christian Gissel[2], PhD; Alher Mauricio Hernández[1], PhD

[1]Bioinstrumentation and Clinical Engineering Research Group, Bioengineering Department, Engineering Faculty, Universidad de Antioquia, Medellín, Colombia

[2]Department of Health Economics, Justus Liebig University Giessen, Giessen, Germany

Corresponding Author:
Alher Mauricio Hernández, PhD
Bioinstrumentation and Clinical Engineering Research Group, Bioengineering Department, Engineering Faculty
Universidad de Antioquia
Calle 70 No. 52-21
Medellín, 050010
Colombia
Phone: 57 2198589
Email: alher.hernandez@udea.edu.co

Abstract

Background: The quest for improved diagnosis and treatment in home health care models has led to the development of wearable medical devices for remote vital signs monitoring. An accurate signal and a high diagnostic yield are critical for the cost-effectiveness of wearable health care monitoring systems and their widespread application in resource-constrained environments. Despite technological advances, the information acquired by these devices can be contaminated by motion artifacts (MA) leading to misdiagnosis or repeated procedures with increases in associated costs. This makes it necessary to develop methods to improve the quality of the signal acquired by these devices.

Objective: We aimed to present a novel method for electrocardiogram (ECG) signal denoising to reduce MA. We aimed to analyze the method's performance and to compare its performance to that of existing approaches.

Methods: We present the novel Redundant denoising Independent Component Analysis method for ECG signal denoising based on the redundant and simultaneous acquisition of ECG signals and movement information, multichannel processing, and performance assessment considering the information contained in the signal waveform. The method is based on data including ECG signals from the patient's chest and back, the acquisition of triaxial movement signals from inertial measurement units, a reference signal synthesized from an autoregressive model, and the separation of interest and noise sources through multichannel independent component analysis.

Results: The proposed method significantly reduced MA, showing better performance and introducing a smaller distortion in the interest signal compared with other methods. Finally, the performance of the proposed method was compared to that of wavelet shrinkage and wavelet independent component analysis through the assessment of signal-to-noise ratio, dynamic time warping, and a proposed index based on the signal waveform evaluation with an ensemble average ECG.

Conclusions: Our novel ECG denoising method is a contribution to converting wearable devices into medical monitoring tools that can be used to support the remote diagnosis and monitoring of cardiovascular diseases. A more accurate signal substantially improves the diagnostic yield of wearable devices. A better yield improves the devices' cost-effectiveness and contributes to their widespread application.

(JMIR Med Inform 2022;10(11):e40826) doi:10.2196/40826

KEYWORDS

signal denoising; motion artifacts; biomedical signal processing; electrocardiogram; ECG; biomedical monitoring; home health care

Introduction

Problem Statement

Digital health provides the opportunity to combat pandemics, to deliver health care in remote regions, and to reduce the carbon footprint of health care delivery. Telehealth and remote monitoring, particularly in patients' homes, has become an important option due to problems associated with keeping patients in hospitals and care centers for extended periods: the increase in the probability of acquiring nosocomial infections [1]; the deficiency in medical infrastructure to meet the demand of patients [2]; the increase in therapeutic dependence by older adult patients [3]; and the increase in hospitalization costs. These issues have led to a search for alternatives in medical care such as home health care.

This has led to an increase in the development of remote monitoring technologies of patients' vital signs to improve the medical diagnosis [4,5]. Wearable devices for monitoring vital signs have become a powerful tool to improve health services and to implement home health care models [6,7]. These will allow to acquire vital signs of patients in daily environments while they carry out their activities in a normal way, allowing to obtain complementary information to improve the medical diagnosis, to constantly monitor the patient's condition, and to improve their treatments [7,8].

A major challenge for the diffusion of digital health technologies lies in the signal quality of sensors for remote diagnosis and monitoring of patients, including electrocardiogram (ECG) signals of moving patients, which require the denoising of motion artifacts (MA). ECG is an important diagnostic technique for application in wearable devices, owing to the amount of information contained in the acquired waveform, distributed in different peaks and undulations called segments (P, Q, R, S, T, and U). The P segment represents atrial depolarization, the QRS complex represents ventricular depolarization and atrial repolarization, the isoelectric ST segment is the time when both ventricles are completely depolarized, the T segment represents ventricular repolarization, and the U segment represents papillary muscle repolarization [9,10]. Each segment is characterized by its unique shape, amplitude, duration, and time of occurrence, allowing to identify the way in which the electrical impulse is conducted through the heart muscle [10].

State of the Art

Newer devices have been developed to identify cardiovascular diseases in early stages (asymptomatic). Some of them are external loop recorders, implantable loop recorders, and Zio patches as well as wearable ECGs. Similar to traditional Holter, they share a sensitivity to artifacts [11], which leads to repeated ECG monitoring with cost increases in 11.1% of cases [9]. In the case of wearable ECGs, the information provided by these devices is not considered for clinical use owing to the contamination by different noise sources such as power line interference, baseline wander, and MA that have nonlinear, nonstationary, and unpredictable character, as well as ECG bandwidth overlaps [12-14]. In addition, the effects of daily life movements on the signal are difficult to predict, which makes the devices' validation for medical use in home health care and outdoor conditions even more difficult. This motivates the development of techniques that allow the reduction of interference in the ECG signal.

Previously, research has been conducted to develop techniques that solve the problem of combined interference of MA, baseline wander, and power line interference in ECG signals [15]. Performance assessments of denoising techniques such as wavelet shrinkage (WS), empirical mode decomposition (EMD), wavelet independent component analysis (WICA), and EMD independent component analysis (ICA) have been performed. These methods present problems, although some of these work in the denoising of synthetic signals when working with signals from patients in movement. One of them is that their signal databases only consider a single source of information (ECG signal) to perform signal denoising, which makes it difficult to acquire the dynamics introduced by movement and significantly affects the performance of artifact reduction methods. However, it was found that depending on the segment of the ECG signal to be preserved, it is possible to use a specific denoising technique for that segment [16].

In recent years, significant advances in the development of techniques for feature extraction from cardiovascular signals in wearable monitoring have been made. The presence of MA has been identified as a significant source of noise in signal acquisition, masking information about the physiological process and leading to misdiagnosis. The MA reduction problem is still addressed in different ways as proposed by Yang and Tavassolian [17], where it is possible to obtain cardiovascular parameters from the seismocardiography signal analysis; they used the ICA on the inertial signals acquired from inertial measurement units (IMUs) and used the ECG and photoplethysmography (PPG) signals as reference signals. An and Stylios [18] evaluated conventional filtering methods using finite impulse response, infinite impulse response, moving average, moving median filters, and advanced decomposition methods such as wavelet, EMD, and adaptive filters to compare their performance. They found that all these methods have their limitations, but the best method was considered to be the adaptive filter. However, it depends on a good selection of the reference signal and still introduces distortion to the signal [18,19].

Other approaches use adaptive noise signal detection, EMD, or wavelet decomposition of the signal of interest and dynamic time warping (DTW) component selection to reduce baseline wandering and high-frequency noise, and they achieved signal improvements of up to 25% [11,20,21]. On the other hand, the electrode configuration and its interaction with the skin had been evaluated to determine the impedance variation and the noise introduction in the signal acquisition [22-24]. Many authors agree that the way to approach the problem is through the use of signal decomposition methods such as wavelets, EMD, and ICA, among others [25-27]. In addition, including multiple sources of information such as pressure signals, PPG and movement are essential to estimate the physiological parameters of interest [28].

Study Objectives

Although some of these methods use the acquisition of multiple signals as other sources of information for the determination of cardiovascular parameters, few of them are focused on evaluating the recovery of the waveform of the ECG signal. Similarly, dealing with the acquisition of multiple physical magnitudes such as pressure or PPG through a single channel does not allow confirmation of the correct denoising of the signal of interest. The method proposed in this paper allows the redundant and simultaneous acquisition of ECG signals and movement signals to obtain more complete information of the physiological process masked by the movement artifacts.

It has been observed that the correct location of the electrodes on the volume conductor of the patient has a great influence on the result of the ECG, to the point that a bad location of these can lead to diagnostic mistakes [29,30]. However, it has been identified that certain modifications in the signal acquisition hardware, such as adding additional leads in the back and taking information of the movement of the person, allow obtaining additional information of the signal of interest [16,31]. In this study, additional electrodes were added to the acquisition hardware of the ECG signal in the chest and back of the volunteers. This represents a novel method for the acquisition of the signals in a redundant and simultaneous way to improve the denoising, decreasing the distortion introduced in the MA reduction process.

The distortion concept in the context of biomedical signal processing refers to the change in the natural shape of the signal due to external processes or disturbances, which include changes in the amplitude, duration, or time of occurrence of the segments that compose the signal and lead to loss of information. The concept of redundant measurement refers to the acquisition of information from the same interest source from ≥2 different measurement points, with the purpose of preventing the loss of information or increasing the sources of information of interest, as is true in this study. Redundant measurements are performed simultaneously to ensure that the information acquired from the different measurement points is synchronized, which is defined as multichannel synchrony.

Further improvement of noise and MA reduction is critical to achieve an optimal diagnostic yield with wearable health care monitoring systems. The diagnostic yield will be an important determinant of the devices' cost-effectiveness [32]. Only with a reliable diagnosis of specific cardiac arrythmias such as atrial fibrillation will the devices' cost-effectiveness allow widespread application, even in resource-constrained environments [33].

Methods

Overview

This paper presents a novel method for the reduction of noise and MA in ECG signals from walking individuals in ambulatory vital signs monitoring applications. We have introduced a new method called Redundant denoising Independent Component Analysis (Rd-ICA). It is based on (1) redundant and simultaneous measurement of ECG signals in the chest and back; (2) acquisition of triaxial movement signals from IMUs; (3) a reference signal synthesized from an autoregressive model, which considers the features of a resting ECG signal obtained through ensemble average (EA) ECG [9,16]; and (4) separation of interest sources and noise sources through multichannel ICA. After the separation of the signals, the identification of the ECG signal is made through the comparison of the components with the synthesized reference ECG signal.

The performance of this method is tested with a database composed of data sets of movement signals and ECG signals acquired in the chest and the back from healthy volunteers in conditions of rest and movement. In addition, the performance of the Rd-ICA method is compared with the performance presented by state-of-the-art denoising methods such as the WS method and the WICA method. The calculation of performance indexes is performed with indexes such as the signal-to-noise ratio (SNR), the DTW, and a proposed index defined as weighted distortion assessment (WDA). It measures the characteristics of the shape of wave found with the EA ECG method.

This section shows the protocol for recording ECG signals, presents some previous state-of-the-art methods for ECG signal denoising, and finally shows the proposed Rd-ICA method. In this work, the comparison of the methods' performance was also carried out. A new index based on the signal distortion characterization through the EA ECG has been proposed.

Register Protocol

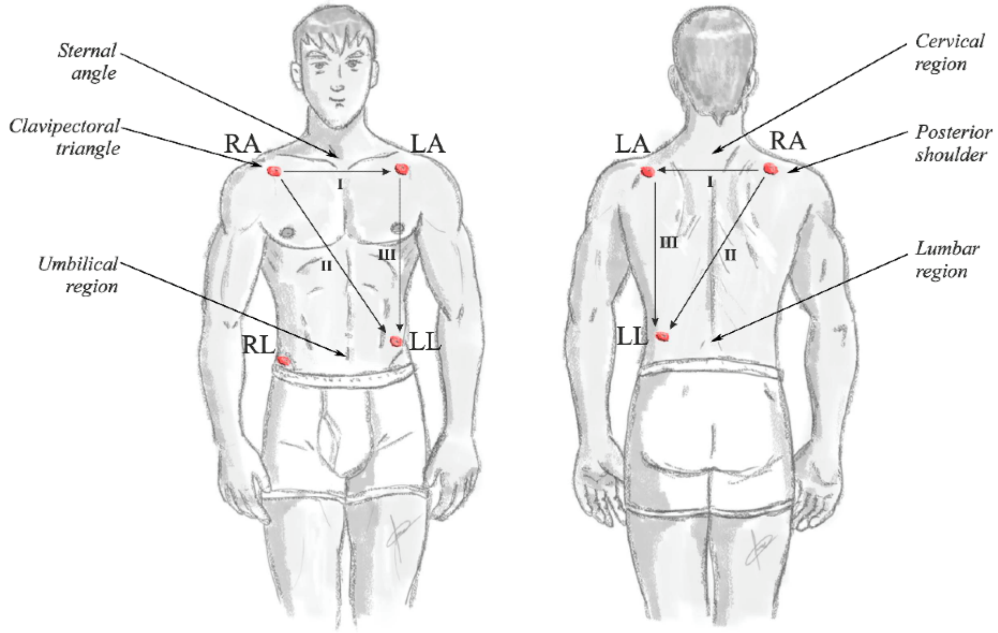
A database of ECG signals acquired from a population of 20 healthy volunteers aged, on average, 26.3 (SD 5.7) years with an average BMI of 24.4 (SD 4.8) kg/m^2 was registered. Database registration was performed for bipolar leads DI, DII, and DIII in the chest and the back and the triaxial movement signal of the volunteer. The experiment was divided into 3 stages: (1) rest before movement, (2) controlled movement in laboratory conditions, and (3) rest after movement. Each volunteer was asked in the first stage to remain at rest for 5 minutes, which led to the acquisition of reference ECG and motion signals. In the second stage, each volunteer was asked to perform a walk at a normal travel speed of 4.2 (SD 0.8) km/h for 5 minutes, which produces contamination that masks and distorts the ECG signal significantly. In the third stage, each volunteer was asked again to be at rest for 5 minutes. This protocol was performed to obtain a database of ECG signals contaminated with MA composed of redundant and simultaneous ECG signals acquired in the chest and the back, also with the movement of the volunteer registered through IMUs [24]. ECG signals were acquired at a sampling frequency of 250 Hz and 24-bit resolution [34]. Informed consent was obtained from all participants involved in the study.

To acquire signals, a custom wearable device was used that performs the acquisition of the signals of ECG; PPG; and noninvasive blood pressure that is redundant, simultaneous, and synchronized [31,35]. The ECG is acquired on the volunteer's chest and back, and an IMU is included on each ECG lead to record the inertial activity and movement. It similarly occurs for the PPG signal that is acquired both in the left wrist and the right wrist and for the noninvasive blood pressure signal that is recorded in both arms through the oscillometric method [8]. The IMUs allow acquiring the inertial activity and movement to analyze the nonlinear dynamics of artifact contamination on the signal.

All the signals are recorded simultaneously and synchronized in time to guarantee signal redundancy and the possibility of applying the Rd-ICA method. For the ECG signal, Ag/AgCl electrodes were used, which have proven to be the ones with the least interference in recording of the signals due to their correct coupling with the skin.

In clinical ECG, it is common to place the electrodes on the arms such as the left and right wrists and the left foot under the Einthoven triangle model. In long-term examinations such as Holters, it is common to use the positioning of the electrodes according to the Einthoven triangle under the Mason-Likar [36] method in which the electrodes are located on the person's chest. This model is the most frequently used in ECG wearable monitoring [36,37]. Signals were acquired through the connection of an electrophysiological signal recording equipment to acquire the biopotentials in the torso and back of 20 volunteers as previously validated [24,31]. A group of sensors were placed in the location proposed by the Einthoven triangle in the Mason-Likar [36] method in the chest. These locations were interpolated on the back of the volunteer, considering anthropometric locations [38]. In addition, the acquisition of triaxial movement signals was performed through an IMU located in the electrophysiological sensors. It should be noted that the acquisition of electrophysiological signals in the chest and back was performed redundantly and simultaneously, synchronized with the movement signals. Figure 1 shows the distribution of electrodes in the chest and back of the volunteer.

Figure 1. Electrodes' distribution for the acquisition of electrocardiogram signals according to the Einthoven triangle in the Mason-Likar [38] method over the chest and back of the volunteer. LA: left arm; LL: left leg; RA: right arm; RL: right leg.



Techniques to Reduce MA

The most frequently used techniques to reduce MA have previously been described in detail, including WS, ICA, and WICA [17].

Wavelet Shrinkage

The discrete wavelet transform (DWT) allows to represent a signal as a set of waves through 2 types of functions called mother wavelet and father wavelet, which contain high and low frequency information [39-41]. DWT is a denoising method for ECG signals in a process known as multiresolution analysis [42], where the signal is decomposed in different levels through Mallat tree decomposition and Daubechies 8 mother wavelet selection [43-45]. Then, the thresholding method reduces the noise components with the RiskShrink algorithm from the signal before the reconstruction through inverse DWT is performed [46,47]. In this study, the WS method was applied on each acquired derivation in the chest of volunteers to perform the denoising of ECG signals to compare the performance with the proposed method [48].

Independent Component Analysis

Some measured signals can be considered a linear mixture of information from independent sources, such as artifacts, noise, and interest signals. It is possible to separate these sources with the ICA method [49,50]. To apply the ICA, it is necessary to have a set of observations x = $(x_1, x_2, \ldots x_m)^T$ taken from *m* sensors. The observations are modeled as the linear combination of a set of signals s = $(s_1, s_2, \ldots s_n)^T$, as is described by the mixing model (equation 1) [49,51].

Where the mixing matrix A = $(a_1, a_2, \ldots a_n)$ that has a size of *m* × *n* and a_i are the vectors of the mixture.

To apply the ICA, it is necessary to assume that the sources' signals are independent; just one component has a Gaussian distribution at most; and the number of sensors must be equal to the number of independent sources ($m = n$) [51]. Through the ICA method, the separation matrix w = $(w_{ij})_{(n \times n)}$ and the n separate signals $Y = (y_1, y_2, \ldots y_n)^T$ can be obtained (equation 2).

$$Y = Wx = WAs = Gs \quad (2)$$

Some of the sources are noise sources and should not be considered in the model to perform the denoising. Then, some elements of the separation matrix W must be forced to 0 to reduce the influence of the artifacts on the interest signal [51].

WICA Denoising

As previously described, some extensions of the ICA using decomposition techniques such as DWT have been proposed to denoise physiological signals [17]. The WICA performs the DWT decomposition to obtain multichannel signals from a single-channel signal before applying the ICA method [52]. The process is outlined in the form of an algorithm in Textbox 1.

In this study, the WICA method was applied to the ECG signals acquired in the chest of the volunteers to evaluate its performance. The selection of noise sources or artifacts is performed conventionally through visual inspection as is proposed by other studies, which is one of the main drawbacks of this method [51,52].

Textbox 1. Wavelet independent component analysis (ICA) algorithm.

Algorithm

1. Select the mother wavelet and the order of the wavelet transform.
2. Apply the wavelet decomposition (discrete wavelet transform [DWT]) to generate the input matrix for the ICA algorithm.
3. Apply the ICA method to the set of wavelet components and derive the corresponding mixing (A) and demixing (W) matrices.
4. Select the sources of interest, force the others to 0, and multiply this selection with the mixing matrix (A) to back-reconstruct their appearance in the set of wavelet components.
5. Apply the inverse DWT over the new set of wavelet components to back-reconstruct the enhanced signal.

EA Electrocardiogram

The EA ECG allows the characterization of the ECG signal through the measurement of segments' features that compose the ECG signal (P, Q, R, S, T, and U). This method has been used to evaluate the ECG signal distortion introduced by MA. It allows to evaluate the performance of denoising methods quantitatively considering the waveform of the signal. This is done by finding the average pattern of a signal that has a periodically repeated waveform, which is the case for the ECG signal [16].

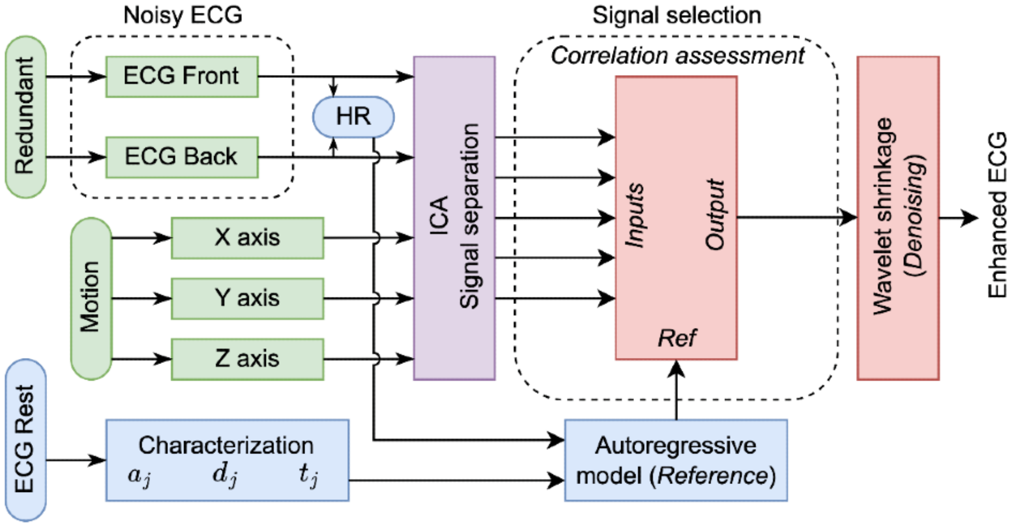
In the EA ECG computing process, it was necessary to select a fiducial point on the standard waveform, which was the reference in time to synchronize the signal waveforms. The R peak was selected because it has the maximum amplitude in waveform. On the other hand, the size in time or in samples that have the standard waveform was determined to perform the partition of the signal, the synchronization through the fiducial points and the averaging of the signals. This time was determined as the time elapsed between 2 consecutive R peaks and corresponds with the heart rate.

In this study, the method was used first to perform the characterization of the ECG signals acquired at rest as a reference. In addition, the method was used to measure the performance by state-of-the-art denoising methods and the proposed Rd-ICA method. Furthermore, the features obtained through the EA ECG were used to synthesize a reference signal from an autoregressive model that considers these features and the heart rate to present a synthetic ECG signal that resembles its real counterpart [9].

Redundant Denoising ICA Method

The method proposed in this paper is based on the simultaneous and redundant acquisition of the ECG signal, the movement of the person, the separation of multichannel components, and the selection of the improved signal through the comparison with a modeled signal from previous information. The processing scheme of the Rd-ICA method is presented in Figure 2.

Figure 2. Block diagram representing the proposed Redundant denoising Independent Component Analysis method. The method considers the simultaneous and redundant acquisition of the electrocardiogram (ECG) signal contaminated with motion artifacts (green), the signal characteristics determination and the reconstruction of a reference ECG signal from a resting ECG signal of the same volunteer (blue), the components separation (purple), and the selection of the improved ECG signal (watermelon). HR: heart rate; ICA: independent component analysis.



Acquisition of Redundant ECG and Motion Signals

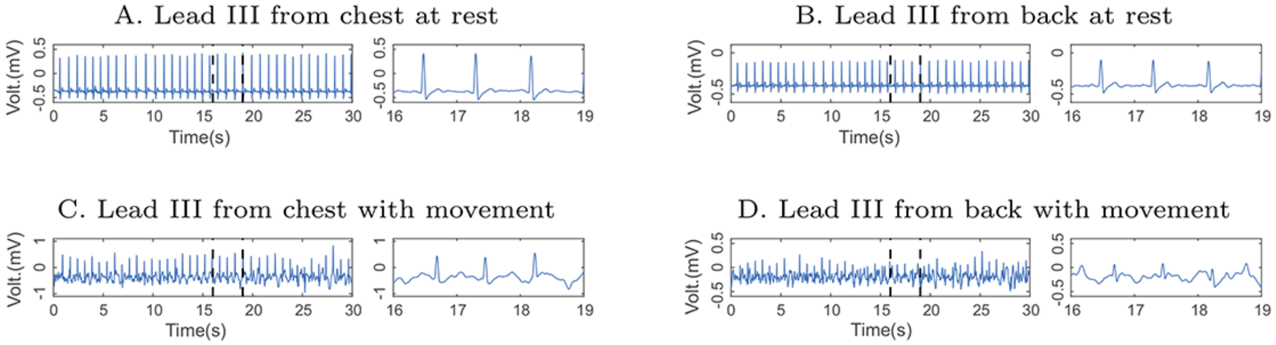
The first step of the method consists in the acquisition of the redundant and simultaneous ECG signals and the movement signals of the person. The acquisition of the ECG signals in leads DI, DII, and DIII was performed both in the chest and back of the volunteers [24]. Redundancy of the signals is achieved by acquiring the leads of the ECG signal in the chest and back of the volunteer [53]. The acquisition in the chest and back is carried out simultaneously so the leads DI, DII, and DIII acquired in both areas are synchronized in time. The volunteers' ECG segments acquired at rest were analyzed by a cardiologist to validate that the volunteers did not present evident cardiac pathology before analysis, confirming that the parameters of the signal segments are within the normal range in physiological terms.

Some of the ECG signals acquired from a healthy volunteer used in this work are presented (Figure 3). Figure 3A and Figure 3B show the ECG signal acquired at rest in the chest and back of the volunteer, respectively. Figure 3C and Figure 3D show the ECG signal with MA acquired in the chest and back of the volunteer, respectively, while the volunteer performed movement. Each figure has a sample of 30 seconds and a detail of 3 seconds to show the waveform. In addition, the synchronization between the signals acquired in the volunteer's chest and back was presented, as there is no lag in the QRS complexes of each pair of signals.

The electrode movement pattern was acquired by adding an IMU on each electrode. That information was acquired simultaneously with the ECG signals, thus increasing the amount of information available for multichannel signal analysis. For the ECG signals contaminated with MA, the time that elapses between 2 consecutive R peaks was measured. This measurement represents the heart rate of the ECG signal during the movement. This feature of the signal was used to synthesize the reference ECG signal.

To calculate the heart rate, it is necessary to measure the time between 2 consecutive heartbeats. It is common to identify the QRS complex and measure the time elapsed between 2 of them consecutively. This method was used to calculate the heart rate, both in the signals acquired at rest and while moving (equation 3).

Figure 3. Epochs of 30-second lead III electrocardiogram (ECG) acquired in the chest and back of healthy volunteers at rest and with motion artifacts. The left panel in 4 axes shows 30-second epochs, while the right one shows a 3-second detail of the ECG signals. (A) Lead III from chest at rest, (B) lead III from back at rest, (C) lead III from chest with movement, and (D) lead III from back with movement.



Modeling the Reference ECG Signal

To perform the comparison and selection of the resulting components after the multichannel analysis, a reference ECG signal was modeled from an autoregressive model that considers the features of the ECG signal segments such as amplitude, duration, and time of appearance of each segment [9,54]. It should be noted that the synthesized signal from an autoregressive model is only used as a comparison reference to determine which of the components obtained in the Rd-ICA method contains the highest percentage of information on the physiological signal of interest.

Synthesis of the reference ECG signal is based on the modeling of a group of functions $S_0(t)$, which represent each of the segments of the ECG signal as a modified waveform of the $S_j(t)$, which is obtained through Fourier models in the time interval ($0 \leq t \leq T$) [9]. This time interval corresponds to the time between 2 consecutive R peaks that the modeled signal must have and is taken from the measured heart rate of the signal contaminated with MA. The mathematical equation describes the construction of each segment of the reference ECG signal (equation 4).

$$S_0(t) = a_j S_j(d_j t + t_j) + c_j;\ a_j>0, d_j>0 \quad (4)$$

Where a_j, d_j, t_j, and c_j are the coefficients of amplitude, duration, time of appearance, and offset of the signal, respectively. If the variation of the baseline is subtracted in the processing step, c_j can be omitted.

The autoregressive model requires the definition of the features of each segment of the signal (P, Q, R, S, T, and U) and the heart rate that the modeled signal will have. The features of amplitude, duration, and time of appearance of each segment were obtained from applying the EA ECG method to a signal previously obtained during the volunteer's rest [16]. The EA ECG method allows to characterize the waveform of the ECG signal acquired at rest and to extract the coefficients for the synthesis of the reference ECG signal (equation 4). This later has the waveform of the ECG signal at rest, which is free of artifacts but includes the heart rate of the ECG contaminated with MA.

Separation and Selection of Multichannel Sources

With the premise that the interest information of the ECG signal comes from an independent source, which is the heart, and the MA come from sources other than this; the redundant measurement of the ECG signal was used to obtain information from a single source in 2 different sensors. Each of these sensors acquires information from artifacts from different sources. We assumed that redundant signals will have a common component that will be the ECG lead that is being measured in the chest and back of the volunteer and will have independent components from different sources of MA.

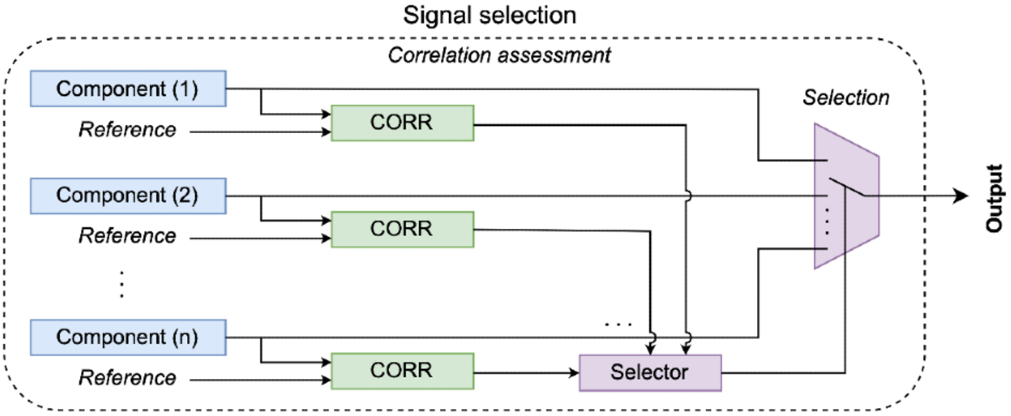
To identify the relative movement of the volume conductor, a set of signals formed by the redundant evaluations of an ECG derivation acquired in the volunteer's chest and back is obtained as well as the movement signals in 3 orthogonal axes obtained with the IMU.

In this regard, the ICA method was used to perform an analysis of the signal data set to identify common information between the different sensed channels and to separate it from independent sources, using the ICA model.

Once the independent components were obtained, the component with more information about the ECG signal was determined. For this, the reference ECG signal was used that was synthesized from the features of the signal acquired at rest from the same volunteer, the heart rate measurement from the ECG signal with MA and the autoregressive model that considers the characteristics and heart rate [9]. Figure 4 presents the assessment and selection method of different components with ECG information.

Each component resulting from the ICA method is compared with the reference signal through the correlation method (Figure 4), which provides a quantitative measure of the similarity between 2 signals. After that, the selection of the component with the highest correlation with the reference ECG signal is made. This component is processed in the final stage of filtering.

Figure 4. Selection method for the component that contains the greatest amount of information of the electrocardiogram signal source. CORR: correlation.



Final Stage of ECG Signal Filtering

Once the component with the most ECG information is obtained, it is filtered and improved with the WS method. With this method, noise filtering of components of the signal with high and low frequency is performed to obtain an improved ECG signal.

Validation and Assessment of the Rd-ICA Method

Overview

To validate the Rd-ICA method, the performance presented by it is evaluated and compared with that of other state-of-the-art denoising methods such as the WS and WICA methods [16].

As presented earlier, the WS and WICA methods were applied only to the leads acquired on the chest of the volunteers, as each of these methods proposes, and the performance presented by these methods was measured. On the other hand, the Rd-ICA method was applied to the same set of signals but included the redundant signals and motion signals as proposed. The performance of this method was also measured.

The performance measurement presented by these methods was performed through the calculation of indexes traditionally used in signal processing. These indexes were the SNR, DTW, cross-correlation, and the measurement of the difference percentage of the signals' features through the EA ECG method with respect to an ECG signal acquired at rest from the same volunteer. For this, an index that considers the distortion introduced by the denoising methods when performing the signal improvement was proposed. This index was called WDA.

Signal-to-Noise Ratio

The SNR was used to quantify the improvement of the enhanced signal after the denoising methods were applied. The SNR reflects the difference between input (reference signal) and output (enhanced signal) of the specific denoising methods (equations 5, 6, and 7) [55,56].

Where x_c is the clean ECG signal, x_n denotes the noisy ECG, and x_d represents the denoised ECG.

Dynamic Time Warping

The DTW method allows to calculate the minimum Euclidean distance between each sample of the signal to be compared $S_j(t)$ and each point of the reference signal $S_0(t)$ [57,58]. The method uses 2 matrices of identical size to perform the calculation. Matrix $S1_{m\times n}$ contains m copies of the reference signal $S_0(t)_{1\times n}$ in the rows, and matrix $S2_{m\times n}$ contains n copies of the signal to compare $S_j(t)_{m\times 1}$ in the columns [17]. The distance matrix $D_{m\times n}$ is calculated using the single dimension Euclidean distance as shown in equation 8.

Where $1 \leq x \leq m$ and $1 \leq y \leq n$. Starting in position (1,1) of D, a cost matrix C is created to store the accumulated distance of the previous column and row, which are calculated with equation 9.

The path of minimum distance is found from cost matrix C, starting at the position (m,n) of the matrix and moving toward the adjacent position of lowest cost until reaching the beginning. These positions are saved and will then be identified in matrices S_1 and S_2 to create the minimum difference aligned signals S_{1w} and S_{2w}, respectively. In this process, it is possible that some samples of the matrix S_1 or S_2 are repeated to conform to the vectors S_{1w} and S_{2w}, which is an index of the difference between both signals and the distortion of the evaluated signal. Through this method, the distance between the standard waveform of the modeled reference ECG signal at rest and the enhanced ECG signal is determined.

Cross-Correlation

Cross-correlation is a measure of similarity between 2 signals. This is measured from the displacement and the superposition of one signal on the other to determine the level of similarity between both. This is known as sliding dot product and is defined in equation 10.

In this work, cross-correlation was used as an index to determine the performance of the denoising methods. Similarity between the reference ECG signal and the enhanced ECG signal was evaluated through the cross-correlation index.

Weighted Distortion Assessment

For the calculation of the WDA index, the percentage similarity vector of the features (△EA) and (△W) and the vector of weighted weights for the features are defined. These 2 vectors are defined by equations 11 and 12.

$$\triangle EA = [a_P, d_P, t_{P\text{-}R}, a_R, d_R, t_{R\text{-}T}, a_T, d_T] \quad (11)$$

$$\triangle W = [w_{P1}, w_{P2}, w_{P3}, w_{R1}, w_{R2}, w_{R3}, w_{T1}, w_{T2}] \quad (12)$$

a_P, d_P, $t_{P\text{-}R}$, a_R, d_R, $t_{R\text{-}T}$, a_T, d_T corresponds to the similarity percentages of each feature obtained from the EA ECG. These are calculated using equation 13 [16]. w_{P1}, w_{P2}, w_{P3}, w_{R1}, w_{R2}, w_{R3}, w_{T1}, w_{T2} correspond to the weights to define the relevance that each feature will have on the calculation of the WDA index. The WDA index is described in equation 14.

$$\text{Similarity} = (100\% - \%\text{difference}) / 100 \quad (13)$$

The coefficients' values of the weight vector (△W) must be chosen between 0 and 1. These represent the percentage of relevance given to the preservation of a certain feature in the WDA performance index assessment. In this study, 4 different cases were evaluated: (1) all features have equal weight, so all the coefficients in the vector (△W) will be equal to 1; (2) the amplitude of the P wave will be more relevant in the analysis, thus w_{P1}=0.9 and the other coefficients will have a weight of 0.5; (3) the amplitude of the QRS complex will be more relevant, w_{R1}=0.9, and the other coefficients will have a weight of 0.5; and (4) the amplitude of the T wave will be more relevant,

w_{T1}=0.9, and the other coefficients will have a weight of 0.5. A high value of the WDA index indicates that the denoising was performed satisfactorily and the distortion introduced by the denoising methods was low, also indicating a better conservation of the features of the ECG signal.

Ethics Approval

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Human Studies Institutional Ethics Committee of Universidad de Antioquia (protocol code 16-59-711 of May 19, 2016). Informed consent was obtained from all participants involved in the study.

Results

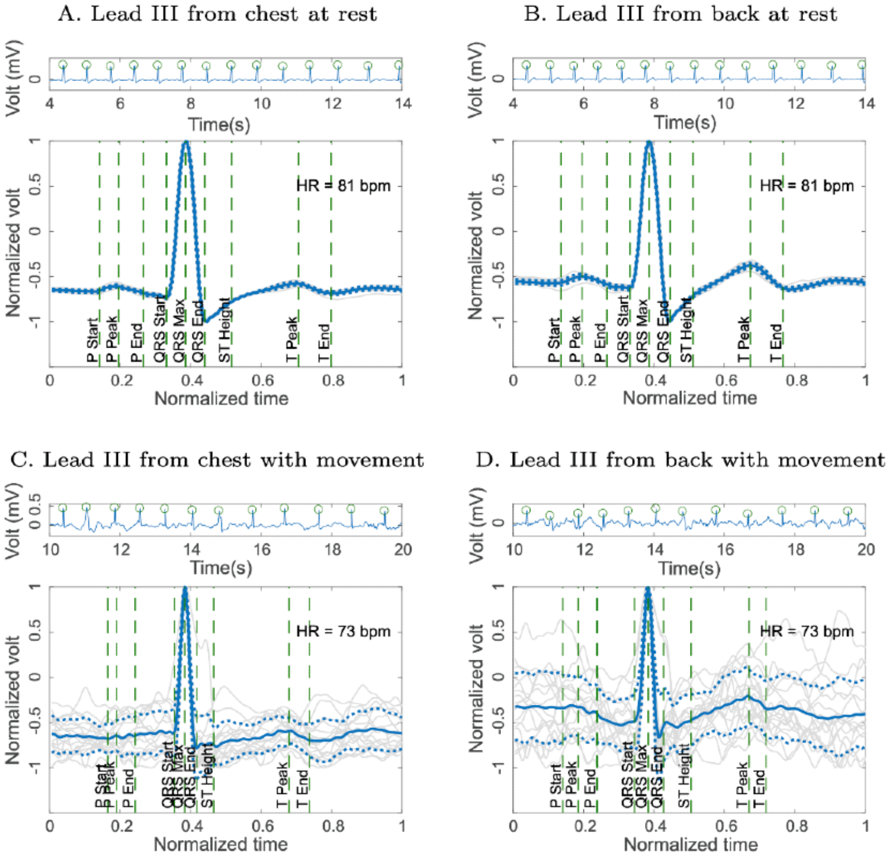
This section shows the results of the Rd-ICA method application for the reduction of MA in ECG signals and the assessment and comparison of the Rd-ICA method performance with the WS and WICA methods applied on ECG signals acquired from healthy volunteers in rest and movement. The performance evaluation of the Rd-ICA method was conducted through the measurement of indexes such as SNR, DTW, and the WDA index proposed in this paper based on the measurement of features from the EA ECG.

EA ECG of the ECG Signal at Rest and Movement

The EA ECG method was applied to ECG signals of volunteers acquired in resting conditions (Figures 3A and 3B) and to ECG signals acquired with MA (Figures 3C and 3D).

Figure 5A and Figure 5B show the EA ECG of the signals acquired in resting conditions in the chest and back of a volunteer, respectively. The solid line represents the average signal and the dashed lines represent the SD signals. Figures 5C and 5D present the acquired signals in movement conditions; in the same way it presents the average signal by means of the continuous line and the SDs by means of the dashed lines. The dispersion of the different waves that compose the EA is observed during the movement.

Figure 5. Ensemble average (EA) electrocardiogram (ECG) for ECG signals acquired at rest and movement, on the chest and back of the volunteer. The continuous centerline represents the EA ECG of the signal while the dashed lines represent the SD of the EA ECG. (A) Lead III from chest at rest, (B) lead III from back at rest, (C) lead III from chest with movement, and (D) lead III from back with movement. HR: heart rate; P: P segment; Q: Q segment; QRS: QRS complex; ST: ST interval; T: T segment.



Application of Denoising Methods

Some state-of-the-art denoising methods were applied to ECG signals contaminated with MA. These signals were selected only from the chest of the volunteers while they performed movement. The methods applied were the WS and WICA method.

The Rd-ICA method was applied to the redundant and simultaneous ECG signals contaminated with MA and acquired in the volunteer's chest and back. The comparison was made with a reference ECG signal synthesized from an autoregressive model. The result of applying denoising methods on the ECG signal of a volunteer is presented (Figure 6). In addition, the EA ECG of the denoising result of the ECG signal through the WS, WICA, and Rd-ICA methods is shown (Figure 7).

A graphic comparison between the results of the denoising presented by each of the evaluated methods is presented (Figure 6). In addition, the result of applying the EA ECG over the improved signals with each of the methods, including average and SD signals is presented (Figure 7).

Figure 6. Epochs of 30-second lead III electrocardiogram (ECG) signal contaminated with motion artifacts (MA) after denoising and enhancement. The left panel in 4 axes shows 30-second epochs, while the right one shows a 4-second detail of the ECG signals. (A) ECG lead III with MA, (B) ECG lead III with WS denoising, (C) ECG lead III with WICA denoising, and (D) ECG lead III with Rd-ICA denoising. Rd-ICA: Redundant denoising Independent Component Analysis; WICA: wavelet independent component analysis; WS: wavelet shrinkage.

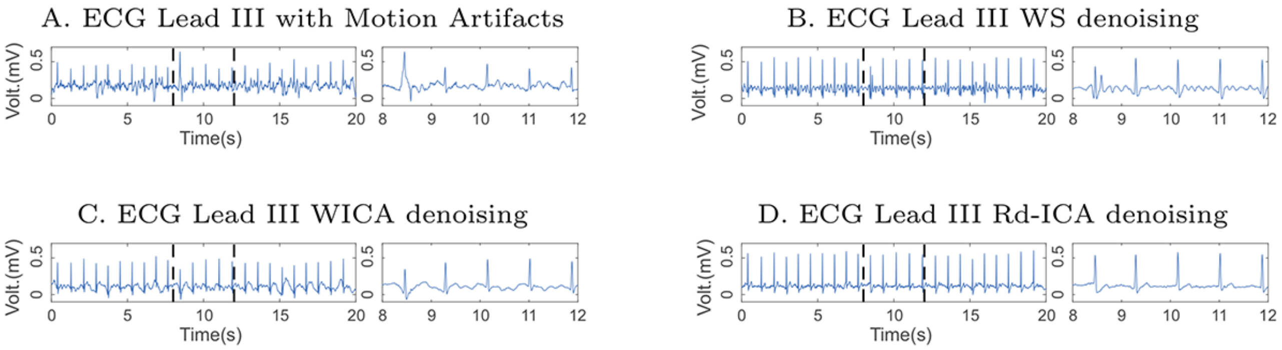
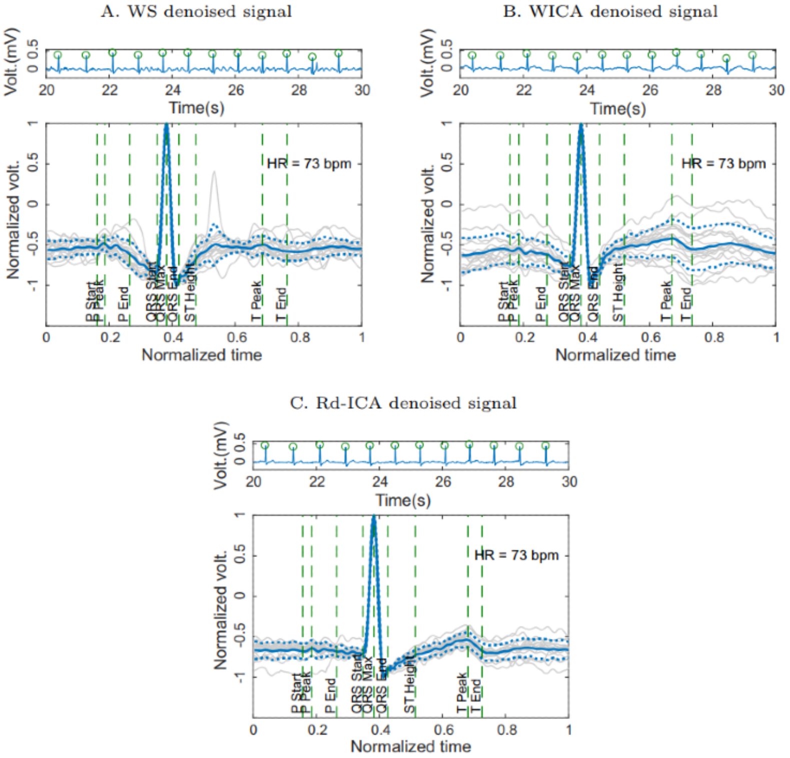


Figure 7. Ensemble average (EA) electrocardiogram (ECG) for enhanced ECG signals through the wavelet shrinkage (WS), wavelet independent component analysis (WICA) and Redundant denoising Independent Component Analysis (Rd-ICA) methods. The continuous centerline represents the EA ECG of the signal while the dashed lines represent the SD. (A) ECG lead III with WS denoising, (B) ECG lead III with WICA denoising, and (C) ECG lead III with Rd-ICA denoising. HR: heart rate; P: P segment; Q: Q segment; QRS: QRS complex; ST: ST interval; T: T segment.



Validation and Assessment of the Rd-ICA Method

SNR Results

To determine the performance of the denoising methods, the improvement SNR (SNR_{imp}) was evaluated on the enhanced ECG signals (Figure 6). The SNR_{imp} was calculated using equation 7 as the difference between the SNR_{in} measured on the ECG signal contaminated with MA before enhancement and the SNR_{out} measured on the enhanced ECG signal through each method. The SNR_{in} obtained for the signal contaminated with MA was −4.64 (SD 0.43). A SNR_{out} of −3.94 (SD 0.35) for WS, −3.22 (SD 0.34) for WICA, and −1.07 (SD 0.15) for Rd-ICA was obtained. The above represents a SNR_{imp} of 0.70 (SD 0.37) for WS, 1.42 (SD 0.31) for WICA, and 3.58 (SD 0.36) for the Rd-ICA method. A larger SNR_{imp} was observed in the Rd-ICA method, followed by the WICA method and finally by the WS method.

DTW Results

Through the DTW method, it was possible to identify the distance percentage between a waveform of a reference signal with the waveform of a signal under evaluation. This percentage was calculated from the DTW method [57]. The lower the percentage, the greater the similarity between the evaluated signal and the reference signal.

The DTW showed 42.62% (SD 8.64%) of difference between MA contaminated signal and the reference signal. Distance percentages of 41.32% (SD 8.01%) for WS, 44.23% (SD 16.31%) for WICA, and 19.72% (SD 6.25%) for Rd-ICA were obtained. A smaller percentage of distance was obtained in the result of the Rd-ICA method. For the WICA and WS methods, an increase in the distance percentage was observed with respect to the signal contaminated with MA, which suggests a distortion increase.

Cross-Correlation

Cross-correlation provides information about the similarity between 2 signals. In this case, the similarity between the reference ECG signals obtained from the model and the enhanced ECG signals was evaluated. For ECG signal contaminated with MA it was obtained a cross-correlation of 6.08 (SD 1.48). Cross-correlation of 5.74 (SD 0.67) for WS, 5.84 (SD 1.31) for WICA, and 8.66 (SD 0.39) for Rd-ICA were obtained.

WDA and Difference Percentage

From the EA ECG, the features of the resting ECG signals, ECG signals contaminated with MA, and the enhanced ECG signals were measured. The difference percentage of the features between the contaminated and enhanced ECG signals relative to the signal acquired at rest was measured [16]. Table 1 shows the difference percentage for each ECG signal feature through the EA ECG method. The average and SD values are presented.

The WDA index assessment was performed for 4 different cases:

1. All features with equal relevance level .
2. Greater relevance of the P wave .
3. Greater relevance of the QRS complex .
4. Greater relevance of the T wave .

Table 2 shows the result of the distortion analysis through the WDA index for signals contaminated with MA and enhanced through denoising methods.

Previous results show the method that presents the best performance in denoising and that preserves the signal waveform features with less distortion is the Rd-ICA method. This supports the results obtained from the other indexes evaluated.

Table 1. Difference percentage between the features of enhanced electrocardiogram signals and electrocardiogram signals acquired at rest.

Difference	Movement, mean (SD)	Wavelet, mean (SD)	WICA[a], mean (SD)	Rd-ICA[b], mean (SD)
Amplitude P	52.56 (23.20)	59.91 (25.06)	42.03 (36.77)	*27.46 (7.37)* [c]
Duration P	25.86 (19.79)	14.83 (6.78)	30.19 (9.49)	*14.59 (7.58)*
Time P-QRS	30.63 (10.62)	26.99 (16.26)	20.56 (11.50)	*13.87 (8.37)*
Amplitude QRS	18.62 (4.87)	4.38 (1.85)	9.76 (8.61)	*3.24 (2.00)*
Duration QRS	11.29 (7.88)	6.37 (5.04)	9.62 (9.33)	*4.89 (2.58)*
Time QRS-T	40.58 (20.12)	31.05 (14.68)	33.70 (15.90)	*20.10 (7.91)*
Amplitude T	47.01 (19.61)	52.24 (43.02)	53.36 (35.22)	*18.63 (5.53)*
Duration T	27.03 (9.86)	12.06 (8.07)	10.73 (7.80)	*11.54 (5.32)*
Average	31.70 (14.49)	25.88 (15.09)	26.24 (16.83)	*13.04 (5.83)*

[a] WICA: wavelet independent component analysis.

[b] Rd-ICA: Redundant denoising Independent Component Analysis.

[c] Italicized values indicate the best performance in the dynamic time warping assessment.

Table 2. Evaluation of the weighted distortion assessment (WDA) index from the enhanced electrocardiogram signals for 4 different cases of specific feature relevance.

WDA	Movement	Wavelet	WICA[a]	Rd-ICA[b]
Case 1	1.93	2.10	2.09	*2.35* [c]
Case 2	1.74	1.85	1.91	*2.18*
Case 3	1.87	2.05	2.02	*2.26*
Case 4	1.76	1.88	1.87	*2.16*

[a] WICA: wavelet independent component analysis.

[b] Rd-ICA: Redundant denoising Independent Component Analysis.

[c] Italicized values indicate the best performance in the dynamic time warping assessment.

Discussion

Principal Findings

This paper presents a novel method for MA reduction in ECG signals through the acquisition of redundant and simultaneous signals, acquisition of the person's movement, modeling of the reference signal from prior knowledge of the ECG signals at rest, and processing of this set of signals through multichannel processing techniques. This method was called Rd-ICA. This was applied to a data set composed of ECG signals acquired from healthy volunteers at rest and in movement conditions.

The performance of the proposed Rd-ICA method was compared with state-of-the-art methods for MA reduction such as WS and WICA. To compare the performance of the different methods, each was applied to the data set, and the performance was evaluated from the measurement of indexes such as SNR, DTW, cross-correlation, and the proposed WDA that consider the morphological features of the signal [16].

The use of a single denoising method does not guarantee the reduction of noise in all features. Sometimes it is necessary to use specific denoising methods to improve the signal and maintain some feature with fidelity. Despite this, the Rd-ICA method presents a good alternative for the reduction of MA in contaminated ECG signals, as shown by the results obtained in this paper.

Comparison With Prior Works

The ECG signals acquired redundantly in the chest and back of the volunteers showed similar information from the same source; this information belongs to the ECG signal that is of interest in this study [59]. In addition, it contains information from other sources among which are MA, which are measured as information from independent sources. This allowed the separation of the signal of interest and the signal of artifacts through the proposed multichannel signal processing technique.

Most of the methods reported are based on the separation by components, the extraction of the noise components and their elimination, then the reconstruction of the signal of interest. The works at the frontier of knowledge make use of methods such as EMD, wavelet, and adaptive filters with some modifications and improvements, and these are the ones that have presented the best performance, but according to those reports, they also introduce a large amount of distortion in the signal of interest.

The method proposed in this paper is advantageous and novel from the point of view that it acquires the signal redundantly, thus providing a way to validate the processing applied to the signal of interest. Other finding was that both components of the ECG signal acquired on chest and back were significantly similar, while the artifacts contamination showed differences between the 2 signals acquired on the chest and back of volunteers [24]. In addition, the motion component helps to determine the dynamics of the signal, evaluate the nonlinearity of contamination by artifacts and perform a better estimation of the components of interest through the proposed Rd-ICA method.

Application Spectrum

This technique presents its application with wearable vital signs monitoring devices, which have their main field of application in outpatient vital signs monitoring. This technique requires a modification in the signal acquisition hardware as it requires redundant and simultaneous ECG signal acquisition from the chest and back of patients. In addition, it requires the measurement of movement through IMUs. These modifications are possible to implement in wearable devices [31]; therefore, the technical feasibility can be affirmed for implementation in outpatient monitoring.

The proposed technique has important potential in the processing of physiological signals from different sources with the use of redundant acquisition of the interest signals. It shows its application in the identification of signals from cardiac arrhythmias in conjunction with the EA ECG method and the proposed WDA index. Some of the ECG monitoring devices that potentially may include this technique to reduce artifacts and to improve their diagnostic potential are external loop recorders, implantable loop recorders, traditional Holter, and wearable ECGs. This kind of improvement will reduce the associated costs with repeated tests.

The method proposed in this paper presents a considerable advance in the reduction of MA in ECG signals as the results showed an improvement in the denoised signal. Its advantage is not only in the evaluation of signal indexes such as SNR but also in the preservation of the signal waveform, low distortion introduction, and the potential use in medical diagnosis. It was observed that the Rd-ICA method presents a significant improvement in the conservation of the characteristics of the signal compared with the WS and WICA when it was evaluated using the EA ECG method.

In the same way, the possibility to have the redundant ECG signal and the motion signals from the inertial sensors provides the proposed method with the capability to separate the component of interest from the components of artifacts. At the same time, it allows the method to determine these components autonomously by comparing this signal with a reference signal built from an autoregressive model that considers the cardiovascular characteristics of the volunteer. This allows the proposed method to be implemented in wearable monitoring systems and in autonomous monitoring systems.

Future Work

On the other hand, there is evidence of the need to carry out more extensive research to evaluate the reliability and clinical validity of this method in patients with relevant diseases. Although this method presents the possibility to differentiate events coming only from the cardiovascular system and separate them from external events such as the volunteer's movement. This is due to the possibility of obtaining the ECG signals redundantly and simultaneously. In addition, the need to add new sensors for the redundant measurement of the signals and the obtaining of the movement signal through IMUs implies significant modifications in the hardware of the existing monitoring systems. Despite this, the possibility of making these modifications is evident and opens the possibility to generate new designs, enabling the growth of the market for wearable medical devices.

Currently, the research and development of biosensors have a great boom thanks to the advantages they show, such as easy application and portability. Their easy implementation in wearable devices for the continuous measurement of different vital sign signals or physiological variables in everyday environments makes these devices a great option in vital signs monitoring [60]. Despite its great applicability, the information acquired by these biosensors is distorted by different sources of noise and artifacts, such as MA. The method proposed in this work is not limited to the denoising of the ECG signal, but it can be used for other physiological signals that can be redundantly acquired and are susceptible to be affected by MA. Some of these sensors are PPG biosensors, enzymatic biosensors for glucose measurement, intraocular pressure, and hydration percentage [60,61].

Conclusions

The technique's ability to improve the quality of the signal is critical for diagnosing specific cardiac arrhythmias in real-world use. The diagnostic yield has been shown to be a major determinant in a technique's economic assessment; for example, in diagnosis after palpitations [62] or syncopes [63,64], in screening of athletes [65,66], or in identifying asymptomatic atrial fibrillation [33,67]. To explore these applications, the acquisition of a database that considers more extreme movements and patients with common cardiac pathologies is required, which will provide information about the effect of artifact promotion techniques in the correct identification of arrhythmias or the malfunction of heart. Such a database would allow future work on the proposed method and a benchmark with existing methods to evaluate their performance in MA reduction as well as its benefits in the identification of waveforms modified by specific cardiac arrhythmias.

Acknowledgments

This work was supported by Convocatoria 727—Programa de Becas de Doctorado Nacionales 2015 de Colciencias, Republic of Colombia.

This research was funded by Minciencias grant 727—Programa de Becas de Doctorado Nacionales 2015 de Colciencias, Republic of Colombia.

Authors' Contributions

FACU and AMH conceived the experiments; FACU, CG, and AMH designed the experiments; FACU performed the experiments; FACU and AMH analyzed the data; FACU wrote the paper with contributions from CG and AMH. All authors have read and agreed to the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Gernant SA, Snyder ME, Jaynes H, Sutherland JM, Zillich AJ. The effectiveness of pharmacist-provided telephonic medication therapy management on emergency department utilization in home health patients. J Pharm Technol 2016 Oct 01;32(5):179-184 [FREE Full text] [doi: 10.1177/8755122516660376] [Medline: 28924623]
2. Price JR, Cole K, Bexley A, Kostiou V, Eyre DW, Golubchik T, Modernising Medical Microbiology informatics group. Transmission of Staphylococcus aureus between health-care workers, the environment, and patients in an intensive care unit: a longitudinal cohort study based on whole-genome sequencing. Lancet Infect Dis 2017 Feb;17(2):207-214 [FREE Full text] [doi: 10.1016/S1473-3099(16)30413-3] [Medline: 27863959]
3. Sourdet S, Lafont C, Rolland Y, Nourhashemi F, Andrieu S, Vellas B. Preventable iatrogenic disability in elderly patients during hospitalization. J Am Med Dir Assoc 2015 Aug 01;16(8):674-681. [doi: 10.1016/j.jamda.2015.03.011] [Medline: 25922117]
4. Galarraga M, Serrano L, Martinez I, de Toledo P, Reynolds M. Telemonitoring systems interoperability challenge: an updated review of the applicability of ISO/IEEE 11073 standards for interoperability in telemonitoring. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:6162-6165. [doi: 10.1109/IEMBS.2007.4353761] [Medline: 18003427]
5. Valdivieso H, Duque L, Pérez S. Dispositivo Interfaz para Telemonitoreo de Signos Vitales en Transporte de Pacientes. SIC 15-283315. Bogotá, Colombia: Superintendencia de Industria y Comercio; 2015.
6. Golubnitschaja O, Kinkorova J, Costigliola V. Predictive, preventive and personalised medicine as the hardcore of 'Horizon 2020': EPMA position paper. EPMA J 2014 Apr 07;5(1):6 [FREE Full text] [doi: 10.1186/1878-5085-5-6] [Medline: 24708704]
7. Baig MM, Gholamhosseini H, Connolly MJ. A comprehensive survey of wearable and wireless ECG monitoring systems for older adults. Med Biol Eng Comput 2013 May;51(5):485-495. [doi: 10.1007/s11517-012-1021-6] [Medline: 23334714]
8. Castaño FA, Hernández AM, Sarmiento CA, Camacho A, Vega C, Lemos JD. Redundant measurement of vital signs in a wearable monitor to overcome movement artifacts in home health care environment. In: Proceedings of the IEEE 7th Latin

American Symposium on Circuits & Systems. 2016 Presented at: LASCAS '16; February 28-March 2, 2016; Florianopolis, Brazil p. 299-302. [doi: 10.1109/lascas.2016.7451069]

9. Castaño FA, Hernández AM. Autoregressive models of electrocardiographic signal contaminated with motion artifacts: benchmark for biomedical signal processing studies. In: Proceedings of the VII Latin American Congress on Biomedical Engineering. 2017 Presented at: CLAIB '16; October 26-28, 2016; Bucaramanga, Colombia p. 437-440. [doi: 10.1007/978-981-10-4086-3_110]
10. Rhoades RA, Bell DR. Medical Physiology: Principles for Clinical Medicine. 3rd edition. Philadelphia, PA, USA: Lippincott Williams and Wilkins; 2009.
11. Malghan PG, Kumar Hota M. Grasshopper optimization algorithm based improved variational mode decomposition technique for muscle artifact removal in ECG using dynamic time warping. Biomed Signal Process Control 2022 Mar;73:103437. [doi: 10.1016/j.bspc.2021.103437]
12. Serteyn A, Vullings R, Meftah M, Bergmans JW. Motion artifacts in capacitive ECG measurements: reducing the combined effect of DC voltages and capacitance changes using an injection signal. IEEE Trans Biomed Eng 2015 Jan;62(1):264-273. [doi: 10.1109/TBME.2014.2348178] [Medline: 25137720]
13. Mishra S, Das D, Kumar R, Sumathi P. A power-line interference canceler based on sliding DFT phase locking scheme for ECG signals. IEEE Trans Instrum Meas 2015 Jan;64(1):132-142. [doi: 10.1109/TIM.2014.2335920]
14. Nathan V, Akkaya I, Jafari R. A particle filter framework for the estimation of heart rate from ECG signals corrupted by motion artifacts. Annu Int Conf IEEE Eng Med Biol Soc 2015;2015:6560-6565. [doi: 10.1109/EMBC.2015.7319896] [Medline: 26737796]
15. Ansari S, Ward K, Najarian K. Epsilon-tube filtering: reduction of high-amplitude motion artifacts from impedance plethysmography signal. IEEE J Biomed Health Inform 2015 Mar;19(2):406-417. [doi: 10.1109/JBHI.2014.2316287] [Medline: 24723634]
16. Castaño FA, Hernández AM, Soto-Romero G. Assessment of artifacts reduction and denoising techniques in electrocardiographic signals using ensemble average-based method. Comput Methods Programs Biomed 2019 Dec;182:105034. [doi: 10.1016/j.cmpb.2019.105034] [Medline: 31454749]
17. Yang C, Tavassolian N. An independent component analysis approach to motion noise cancelation of cardio-mechanical signals. IEEE Trans Biomed Eng 2019 Mar;66(3):784-793. [doi: 10.1109/TBME.2018.2856700] [Medline: 30028685]
18. An X, Stylios GK. Comparison of motion artefact reduction methods and the implementation of adaptive motion artefact reduction in wearable electrocardiogram monitoring. Sensors (Basel) 2020 Mar 07;20(5):1468 [FREE Full text] [doi: 10.3390/s20051468] [Medline: 32155984]
19. Yang YS, Lee WC, Ke TC, Wei CP, Lee CC. Adaptive reduction of motion artefact in wireless physiological monitoring microsystems. In: Proceedings of the 3rd International Conference on Sensing Technology. 2008 Presented at: ICSENST '08; November 30-December 3, 2008; Tainan, Taiwan p. 523-526. [doi: 10.1109/icsenst.2008.4757161]
20. Xie X, Liu H, Shu M, Zhu Q, Huang A, Kong X, et al. A multi-stage denoising framework for ambulatory ECG signal based on domain knowledge and motion artifact detection. Future Gener Comput Syst 2021 Mar;116:103-116. [doi: 10.1016/j.future.2020.10.024]
21. Madan P, Singh V, Singh DP, Diwakar M, Kishor A. Denoising of ECG signals using weighted stationary wavelet total variation. Biomed Signal Process Control 2022 Mar;73:103478. [doi: 10.1016/j.bspc.2021.103478]
22. Gan Y, Rahajandraibe W, Vauche R, Ravelo B, Lorriere N, Bouchakour R. A new method to reduce motion artifact in electrocardiogram based on an innovative skin-electrode impedance model. Biomed Signal Process Control 2022 Jul;76:103640. [doi: 10.1016/j.bspc.2022.103640]
23. Yoon S, Lee S, Yun Y, Lee M. A development of motion artifacts reduction algorithm for ECG signal in textile wearable sensor. In: Proceedings of the World Congress of Medical Physics and Biomedical Engineering. 2006 Presented at: IFMBE '06; August 27-September 1, 2006; Seoul, South Korea p. 1210-1213. [doi: 10.1007/978-3-540-36841-0_292]
24. Castaño FA, Hernández AM. Sensitivity and adjustment model of electrocardiographic signal distortion based on the electrodes' location and motion artifacts reduction for wearable monitoring applications. Sensors (Basel) 2021 Jul 15;21(14):4822 [FREE Full text] [doi: 10.3390/s21144822] [Medline: 34300562]
25. Nagai S, Anzai D, Wang J. Motion artifact removal for wearable ECG using stationary wavelet multi-resolution analysis. In: Proceedings of the IEEE 5th International Symposium on Electromagnetic Compatibility. 2017 Presented at: EMC-Beijing '17; October 28-31, 2017; Beijing, China p. 1-5. [doi: 10.1109/emc-b.2017.8260359]
26. Lilienthal J, Dargie W. Extraction of motion artifacts from the measurements of a wireless electrocardiogram using tensor decomposition. In: Proceedings of the 22th International Conference on Information Fusion. 2019 Presented at: FUSION '19; July 2-5, 2019; Ottawa, ON, Canada p. 1-8. [doi: 10.23919/fusion43075.2019.9011290]
27. Wu CC, Chen IW, Fang WC. An implementation of motion artifacts elimination for PPG signal processing based on recursive least squares adaptive filter. In: Proceedings of the 2017 IEEE Biomedical Circuits and Systems Conference. 2017 Presented at: BioCAS '17; October 19-21, 2017; Turin, Italy p. 1-4. [doi: 10.1109/biocas.2017.8325141]
28. Aygun A, Jafari R. Robust heart rate variability and interbeat interval detection algorithm in the presence of motion artifacts. In: Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics. 2019 Presented at: BHI '19; May 19-22, 2019; Chicago, IL, USA p. 11-15. [doi: 10.1109/bhi.2019.8834543]

29. Malmivuo J, Suihko V, Eskola H. Sensitivity distributions of EEG and MEG measurements. IEEE Trans Biomed Eng 1997 Mar;44(3):196-208. [doi: 10.1109/10.554766] [Medline: 9216133]
30. Rush S, Driscoll DA. EEG electrode sensitivity--an application of reciprocity. IEEE Trans Biomed Eng 1969 Jan;16(1):15-22. [doi: 10.1109/tbme.1969.4502598] [Medline: 5775600]
31. Castaño Usuga FA, Hernández Valdivieso AM. Monitor de signos vitales vestible con interconexión. Google Patents. 2015. URL: https://patents.google.com/patent/WO2017089986A1/es [accessed 2022-11-02]
32. Patel UK, Malik P, Patel N, Patel P, Mehta N, Urhoghide E, et al. Newer diagnostic and cost-effective ways to identify asymptomatic atrial fibrillation for the prevention of stroke. Cureus 2021 Jan 02;13(1):e12437 [FREE Full text] [doi: 10.7759/cureus.12437] [Medline: 33552757]
33. Evans GF, Shirk A, Muturi P, Soliman EZ. Feasibility of using mobile ECG recording technology to detect atrial fibrillation in low-resource settings. Glob Heart 2017 Dec;12(4):285-289 [FREE Full text] [doi: 10.1016/j.gheart.2016.12.003] [Medline: 28302547]
34. Luo S, Johnston P. A review of electrocardiogram filtering. J Electrocardiol 2010;43(6):486-496. [doi: 10.1016/j.jelectrocard.2010.07.007] [Medline: 20851409]
35. Hernández A, Camacho A, Castaño F, Sarmiento C. Movitals. Dirección Nacional de Derecho de Autor. 2017. URL: http://derechodeautor.gov.co:8080/registro-de-software [accessed 2022-11-01]
36. Mason RE, Likar I. A new system of multiple-lead exercise electrocardiography. Am Heart J 1966 Feb;71(2):196-205. [doi: 10.1016/0002-8703(66)90182-7] [Medline: 5902099]
37. Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology, American College of Cardiology Foundation, Heart Rhythm Society, et al. Recommendations for the standardization and interpretation of the electrocardiogram: part I: the electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Computerized Electrocardiology. Circulation 2007 Mar 13;115(10):1306-1324. [doi: 10.1161/CIRCULATIONAHA.106.180200] [Medline: 17322457]
38. Netter FH. Atlas of Human Anatomy. 6th edition. Amsterdam, The Netherlands: Elsevier; 2014.
39. Daubechies I. Ten Lectures on Wavelets (CBMS-NSF Regional Conference Series in Applied Mathematics). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics; 1992.
40. Fehér Á. Denoising ECG signals by applying discrete wavelet transform. In: Proceedings of the 2017 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM) & 2017 Intl Aegean Conference on Electrical Machines and Power Electronics. 2017 Presented at: ACEMP '17; May 25-27, 2017; Brasov, Romania p. 863-868. [doi: 10.1109/optim.2017.7975078]
41. Mallat S. A Wavelet Tour of Signal Processing: The Sparse Way. 3rd edition. Amsterdam, The Netherlands: Elsevier; 2009.
42. Sörnmo L, Laguna P. Bioelectrical Signal Processing in Cardiac and Neurological Applications. Cambridge, MA, USA: Elsevier Academic Press; 2005.
43. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Machine Intell 1989 Jul;11(7):674-693. [doi: 10.1109/34.192463]
44. Kumar A, Singh M. Optimal selection of wavelet function and decomposition level for removal of ECG signal artifacts. J Med Imaging Hlth Inform 2015 Feb 01;5(1):138-146. [doi: 10.1166/jmihi.2015.1369]
45. Singh BN, Tiwari AK. Optimal selection of wavelet basis function applied to ECG signal denoising. Digital Signal Process 2006 May;16(3):275-287. [doi: 10.1016/j.dsp.2005.12.003]
46. Donoho DL. De-noising by soft-thresholding. IEEE Trans Inform Theory 1995 May;41(3):613-627. [doi: 10.1109/18.382009]
47. Taswell C. The what, how, and why of wavelet shrinkage denoising. Comput Sci Eng 2000;2(3):12-19. [doi: 10.1109/5992.841791]
48. Zhang D. Wavelet approach for ECG baseline wander correction and noise reduction. Conf Proc IEEE Eng Med Biol Soc 2005;2005:1212-1215. [doi: 10.1109/IEMBS.2005.1616642] [Medline: 17282411]
49. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw 2000;13(4-5):411-430. [doi: 10.1016/s0893-6080(00)00026-5] [Medline: 10946390]
50. Phegade M, Mukherji P. ICA based ECG signal denoising. In: Proceedings of the 2013 International Conference on Advances in Computing, Communications and Informatics. 2013 Presented at: ICACCI '13; August 22-25, 2013; Mysore, India p. 1675-1680. [doi: 10.1109/icacci.2013.6637433]
51. Xing H, Hou J. A noise elimination method for ECG signals. In: Proceedings of the 3rd International Conference on Bioinformatics and Biomedical Engineering. 2009 Presented at: ICBBE '09; June 11-13, 2009; Beijing, China p. 1-3. [doi: 10.1109/icbbe.2009.5162206]
52. Mijović B, De Vos M, Gligorijević I, Taelman J, Van Huffel S. Source separation from single-channel recordings by combining empirical-mode decomposition and independent component analysis. IEEE Trans Biomed Eng 2010 Sep;57(9):2188-2196. [doi: 10.1109/TBME.2010.2051440] [Medline: 20542760]

53. Malmivuo J, Plonsey R. Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields. Oxford, UK: Oxford University Press; 1995.
54. Kania M, Rix H, Fereniec M, Zavala-Fernandez H, Janusek D, Mroczka T, et al. The effect of precordial lead displacement on ECG morphology. Med Biol Eng Comput 2014 Feb;52(2):109-119 [FREE Full text] [doi: 10.1007/s11517-013-1115-9] [Medline: 24142562]
55. Yan J, Lu Y, Liu J, Wu X, Xu Y. Self-adaptive model-based ECG denoising using features extracted by mean shift algorithm. Biomed Signal Process Control 2010 Apr;5(2):103-113. [doi: 10.1016/j.bspc.2010.01.003]
56. Singh G, Kaur G, Kumar V. ECG denoising using adaptive selection of IMFs through EMD and EEMD. In: Proceedings of the 2014 International Conference on Data Science & Engineering. 2014 Presented at: ICDSE '14; August 26-28, 2014; Kochi, India p. 228-231. [doi: 10.1109/icdse.2014.6974643]
57. Varatharajan R, Manogaran G, Priyan MK, Sundarasekar R. Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. Cluster Comput 2017 Jun 22;21(1):681-690. [doi: 10.1007/s10586-017-0977-2]
58. Shorten GP, Burke MJ. Use of dynamic time warping for accurate ECG signal timing characterization. J Med Eng Technol 2014 May;38(4):188-201. [doi: 10.3109/03091902.2014.902514] [Medline: 24758391]
59. Madhav KV, Raghuram M, Krishna EH, Komalla NR, Reddy KA. Extraction of respiratory activity from ECG and PPG signals using vector autoregressive model. In: Proceedings of the 2012 IEEE International Symposium on Medical Measurements and Applications Proceedings. 2012 Presented at: MeMeA '12; May 18-19, 2012; Budapest, Hungary p. 1-4. [doi: 10.1109/memea.2012.6226650]
60. Sonawane A, Manickam P, Bhansali S. Stability of enzymatic biosensors for wearable applications. IEEE Rev Biomed Eng 2017 May 19;10:174-186. [doi: 10.1109/rbme.2017.2706661]
61. Liang Y, Chen Z, Ward R, Elgendi M. Hypertension assessment via ECG and PPG signals: an evaluation using MIMIC database. Diagnostics (Basel) 2018 Sep 10;8(3):65 [FREE Full text] [doi: 10.3390/diagnostics8030065] [Medline: 30201887]
62. Francisco-Pascual J, Santos-Ortega A, Roca-Luque I, Rivas-Gándara N, Pérez-Rodón J, Milà-Pascual L, et al. Diagnostic yield and economic assessment of a diagnostic protocol with systematic use of an external loop recorder for patients with palpitations. Rev Esp Cardiol 2019 Jun;72(6):473-478. [doi: 10.1016/j.recesp.2018.04.013]
63. Reed MJ, Grubb NR, Lang CC, Gray AJ, Simpson K, MacRaild A, et al. Diagnostic yield of an ambulatory patch monitor in patients with unexplained syncope after initial evaluation in the emergency department: the PATCH-ED study. Emerg Med J 2018 Aug;35(8):477-485. [doi: 10.1136/emermed-2018-207570] [Medline: 29921622]
64. Steinberg LA, Knilans TK. Syncope in children: diagnostic tests have a high cost and low yield. J Pediatr 2005 Mar;146(3):355-358. [doi: 10.1016/j.jpeds.2004.10.039] [Medline: 15756219]
65. Vessella T, Zorzi A, Merlo L, Pegoraro C, Giorgiano F, Trevisanato M, et al. The Italian preparticipation evaluation programme: diagnostic yield, rate of disqualification and cost analysis. Br J Sports Med 2020 Feb;54(4):231-237 [FREE Full text] [doi: 10.1136/bjsports-2018-100293] [Medline: 31315826]
66. Albiński M, Saubade M, Menafoglio A, Meyer P, Capelli B, Perrin T, et al. Diagnostic yield and cost analysis of electrocardiographic screening in Swiss paediatric athletes. J Sci Med Sport 2022 Apr;25(4):281-286 [FREE Full text] [doi: 10.1016/j.jsams.2021.11.039] [Medline: 34895837]
67. Arnold RJ, Layton A. Cost analysis and clinical outcomes of ambulatory care monitoring in Medicare patients: describing the diagnostic odyssey. J Heal Econ Outcomes Res 2015 Feb 11;2(2):161-169. [doi: 10.36469/9897]

Abbreviations

DTW: dynamic time warping
DWT: discrete wavelet transform
EA: ensemble average
ECG: electrocardiogram
EMD: empirical mode decomposition
ICA: independent component analysis
IMU: inertial measurement unit
MA: motion artifacts
PPG: photoplethysmography
Rd-ICA: Redundant denoising Independent Component Analysis
SNR: signal-to-noise ratio
WDA: weighted distortion assessment
WICA: wavelet independent component analysis
WS: wavelet shrinkage

Edited by C Lovis, J Hefner; submitted 13.07.22; peer-reviewed by NS Haider, C Baxter; comments to author 08.08.22; revised version received 01.10.22; accepted 22.10.22; published 25.11.22.

Please cite as:
Castaño Usuga FA, Gissel C, Hernández AM
Motion Artifact Reduction in Electrocardiogram Signals Through a Redundant Denoising Independent Component Analysis Method for Wearable Health Care Monitoring Systems: Algorithm Development and Validation
JMIR Med Inform 2022;10(11):e40826
URL: https://medinform.jmir.org/2022/11/e40826
doi:10.2196/40826
PMID:36274196

Original Paper

The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study

Shengjie Dong[1,2], MPH; Chenshu Shi[3], MSc; Wu Zeng[4], PhD; Zhiying Jia[1,5], MPH; Minye Dong[1], PhD; Yuyin Xiao[1], MPH; Guohong Li[1,6], PhD

[1]School of Public Health, Shanghai Jiao Tong University, Shanghai, China

[2]Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China

[3]Center for Health Technology Assessment, China Hospital Development Institute, Shanghai Jiao Tong University, Shanghai, China

[4]Department of International Health, Georgetown University, Washington, DC, United States

[5]Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

[6]China Hospital Development Institute, Shanghai Jiao Tong University, Shanghai, China

Corresponding Author:
Guohong Li, PhD
School of Public Health
Shanghai Jiao Tong University
No.227 South Chongqing Road, Huangpu District
Shanghai, 200025
China
Phone: 86 21 63846590
Email: guohongli@sjtu.edu.cn

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2022/12/e44374/

Abstract

Background: Studies have shown that hospitals or physicians with multiple malpractice claims are more likely to be involved in new claims. This finding indicates that medical malpractice may be clustered by institutions.

Objective: We aimed to identify the underlying mechanisms of medical malpractice that, in the long term, may contribute to developing interventions to reduce future claims and patient harm.

Methods: This study extracted the semantic network in 6610 medical litigation records (unstructured data) obtained from a public judicial database in China. They represented the most serious cases of malpractice in the country. The medical malpractice network of China was presented as a knowledge graph based on the complex network theory; it uses the International Classification of Patient Safety from the World Health Organization as a reference.

Results: We found that the medical malpractice network of China was a scale-free network—the occurrence of medical malpractice in litigation cases was not random, but traceable. The results of the hub nodes revealed that orthopedics, obstetrics and gynecology, and the emergency department were the 3 most frequent specialties that incurred malpractice; inadequate informed consent work constituted the most errors. Nontechnical errors (eg, inadequate informed consent) showed a higher centrality than technical errors.

Conclusions: Hospitals and medical boards could apply our approach to detect hub nodes that are likely to benefit from interventions; doing so could effectively control medical risks.

(JMIR Med Inform 2022;10(11):e35709) doi:10.2196/35709

KEYWORDS

medical malpractice; complex network; scale-free network; hub nodes; patient safety management; health systems

Introduction

Background

Medical malpractice is a complex issue involving many different elements and their mutual relationships. The interacting elements in medical malpractice could comprise individuals (such as physicians and patients) and institutions (such as hospitals). These elements play particular roles in medical malpractice and have strong or weak connections with it. For example, physicians with poor malpractice records are more likely to stop practicing medicine, switch to smaller practice settings [1,2], or practice defensive medicine. Most malpractice cases are brought against the same physician and occur in the same specialty [3-5]. Owing to the complexity of the topic, it is difficult to describe the organizational themes in medical malpractice using a model or mathematical formula.

The construction and structure of networks may help to understand the complex issues—network thinking focuses on relationships among entities rather than on the entities themselves. Network thinking provides novel ways to address difficult problems such as how to control epidemics; how to target diseases that affect complex networks in the body; and, more generally, what kind of resilience and vulnerabilities are intrinsic to natural, social, and technological networks as well as how to exploit and protect such systems [6]. Similarly, establishing a reasonable medical malpractice network is of great significance for examining common patterns among entities. For example, AIDS network studies [7-9] have suggested that safe sex campaigns, vaccinations, and other interventions should be mainly targeted at hubs in sex contact networks. With complex networks and limited resources, hub targeting would be the most cost-effective strategy [10,11].

Medical malpractice in China is an issue that needs immediate attention. According to statistics from the Supreme Court of China [12], there are >10,000 medical lawsuits each year, and the number of cases has increased markedly. The impact of medical litigation cases is excessive. Wang et al [13] and Li et al [14] estimated that approximately 70% of medical lawsuits in China were related to alleged inadequacies in the quality of health care. However, in Denmark and Sweden, medical litigation cases resulting from insufficient quality of care accounted for only approximately 50% of medical lawsuits [15,16]. The frequent occurrence of such cases will not ease the current tense physician-patient relationship [17,18] and could induce defensive medical behavior. It is believed that defensive medicine either promotes the rise of medical costs or reduces care quality. Unlike the soaring insurance costs caused by the “malpractice crisis” in Europe and the United States, the cost to China’s insurance system appears to be stable, but there may be a huge impending crisis. In China, health care services are mainly provided by public hospitals. Hospitals generally do not purchase commercial insurance and, thus, they bear the medical risks. The lack of a medical risk-sharing mechanism makes it more likely that payments incurred by lawsuits will be potentially diverted from patients’ medical costs; this will make the direct and indirect costs of malpractice more difficult to control.

Studies on medical malpractice have mostly investigated what motivates patients to sue and how malpractice claims affect physicians’ behavior—the aim has been to determine the incentives to practice defensive medicine and change treatment patterns. However, the analytic methods of such studies have been limited to describing characteristics, time trends, and associations; each method has had potential drawbacks and limitations. The complex network theory can provide methodological support for understanding the complexity of a health care system; however, few studies have focused on interactive behavior in medical malpractice in terms of network thinking.

Background Literature

In 2000, the US Institute of Medicine released a report titled “To Err Is Human: Building a Safer Health System” [19], which attracted public attention to incidents of medical malpractice. In recent years, in the United States and Europe, there has been an increase in the number of malpractice claims against health care providers as well as in the amount of payment awarded to plaintiffs. Many descriptive studies have undertaken retrospective analyses of claims [20] with respect to specialties [4,19,21], regional factors [13,22], and medical errors [14,19,23]. On this basis, correlation studies have been conducted, including the following areas: correlations between physician traits and claims [24-26], quality of care and claims [27,28], and medical insurance costs and the medical liability system [29,30]. In the United States in particular, researchers have attempted to explain the sudden increases in claims and sharp rise in insurance premium rates; some believe that such trends may have been caused by a decline in care quality or a lack of efficient incentive schemes provided by legislation. Many studies on medical malpractice have examined the characteristics of liability systems and their ability to prevent negligence and make policy recommendations for ongoing system reform. Other studies have focused on analyzing the impact of medical malpractice on physicians’ behavior and their motives for defensive medicine [31-33].

Health care is complex. Renkema et al [33] identified the complexity of care as a major factor affecting the relationship between malpractice claim risk and physicians’ behavior. Given the complexity of health care, complex theory has been applied to studies on health in many ways. A much-cited article in *The British Medical Journal* by Plsek and Greenhalgh [34] has provided a powerful impetus for the application of complex theory in the field of health. This introductory article argued that, to cope with the growing complexity of health care, linear models had to be abandoned and unpredictability accepted, calling for consideration of the complexity of health services. As an emerging field in complexity research, complex network theory abstracts complex systems into networks, with nodes and connected edges to analyze topology and common patterns for systems. Two well-studied models in complex network theory are the small-world network and scale-free network models [10,35,36]. Originally described in social networks, the small-world property means that the distance between any 2 nodes in a network is unexpectedly small. The scale-free network property means that numerous weakly connected nodes (noninfluential nodes) coexist with a few highly connected nodes (influential nodes).

The complex network theory has been used for studies in evaluating health policy, the spread of infectious diseases, and the mechanism of physiological systems. Yue et al [37] investigated the implementation process of essential drug policy in 3 rural areas in China through the lens of complexity. The authors identified the importance of adaptiveness and self-organizing behavior as well as the role of nonlinear feedback loops in the implementation process. In 2001, a research team of sociologists and physicists from Sweden found that the network of human sexual contacts showed a scale-free structure [7]. Other research has drawn similar conclusions. These findings have provided valuable information for epidemic control, such as with AIDS—in the case of limited resources, it is most cost-effective to prioritize behavioral education or vaccination of the hub node (the most influential node) in the sex network. Several studies in brain science have found that human and other animals' brain structures and functional networks have the following features: small-world topologies [38-41], highly connected hub nodes [42], and modular partitions [43]. There has been limited research on applying complex network theory to medical malpractice. This study used data on medical litigation from China and applied the complex network theory aiming to construct the topology of a medical malpractice network.

Methods

Overview

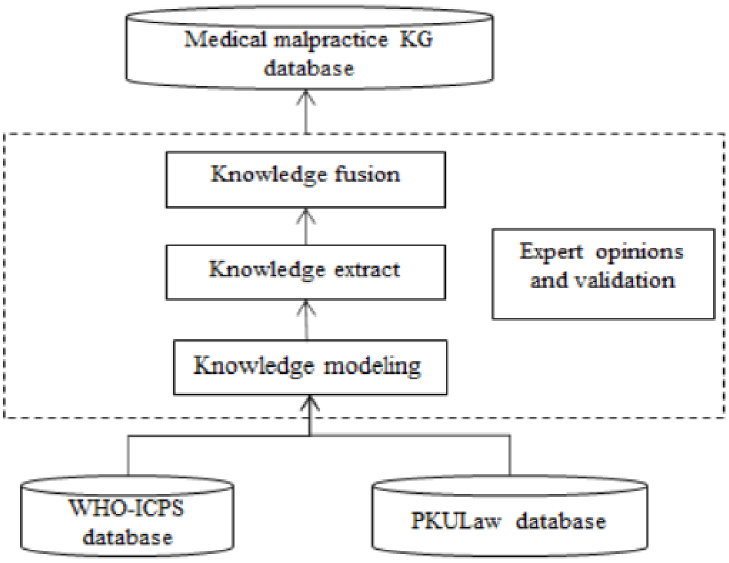
In this study, we constructed a knowledge graph (KG) to represent the medical malpractice network of China (MMNC). Our null hypothesis was that claims are random events—attributable to bad luck with random frequency. Correspondingly, our alternative hypothesis was that medical malpractice is not random; this reflects the belief that hospitals or physicians with multiple malpractice claims are more likely to be involved in new claims. As medical malpractice is a complex issue, this study applied the complex network theory, which provided the methodological support for understanding interactive behavior in medical malpractice. Specifically, this study extracted the semantic network in 6610 medical litigation records (unstructured data) obtained from a public judicial database in China. They represented the most serious cases of malpractice in the country. The MMNC was presented as a KG; it uses the International Classification of Patient Safety from the World Health Organization (WHO-ICPS) as a reference.

Construction of the Malpractice Network

Overview

A complex network can be represented as a KG, which is widely used to express a semantic network. A difficulty in this regard is how to generate an effective, reliable KG. This study followed the general steps of KG development shown in Figure 1. In that process, this research adopted top-down logic (ie, designing the data model first; filling the specific data to the model; and, finally, forming a KG). We stored the KG in Neo4j Community Edition (version 3.5.5; Neo4j, Inc) [44], which is the world's leading graph database and has been widely used because of its higher performance. The structural medical malpractice network can be represented as a KG through the following 4 steps.

Figure 1. Process of constructing the medical malpractice knowledge graph (KG). The International Classification of Patient Safety from the World Health Organization (WHO-ICPS) is a conceptual framework with an ontological basis. However, the WHO-ICPS was not a complete classification at that time. We adopted and localized several key concepts from the WHO-ICPS (details in Table 1).



Step 1: Knowledge Modeling

The knowledge model is the top-level design with a KG—it determines the range of data collected and the structure of the data. From a technological perspective, it defines the *schema* of the KG. In this study, we examined dynamic development in the MMNC—we attempted to determine the underlying mechanism, and the logic can be summarized in chronological order as follows. Patients seek medical advice because of illness. In the case of several medical errors or relatively unsatisfactory outcomes, patients become discontented with the efforts of medical providers and have the incentive to undertake legal action. Each malpractice claim concludes with a legal judgment. The patient, medical provider, and court were considered as stakeholders in the MMNC.

To extract medical litigation texts from the database in China, we referred to the WHO-ICPS [41,45], which offers a conceptual framework using an ontological basis. All definitions and the knowledge model were clarified after repeated discussions by an expert panel (details are described in step 4). The WHO-ICPS is an internationally standardized domain ontology, and it can be directly used as a model when constructing a KG for patient safety. Therefore, this study examined the WHO-ICPS to help construct a theoretical model in step 1.

The actual practice knowledge modeling adhered to the following steps.

First, we defined the network nodes and their properties. The aforementioned stakeholders were classified and served as the nodes. Furthermore, nodes were assigned several properties that were used to form a comprehensive description of the nodes.

Second, we estimated a continuous measure of the association among the nodes. Given that medical litigation cases were in text format, specific sentences that described the relationships among key concepts were abstracted as the relationships. For example, we abstracted “seek medical service” as a relationship from the sentence “Patient A sought medical service from the oncology department at Hospital B.”

Third, we generated an association matrix by compiling all pairwise associations among nodes. We kept the relationships directed (in chronological order) to allow us centrality analysis and weighted (weight was the number of relationships).

Step 2: Knowledge Extraction

From step 2, we obtained information about the nodes and their relationships and properties. Records relating to medical lawsuits were in the form of unstructured data; they covered the contents of patients’ medical records, medical expert opinions, and court decisions. To extract knowledge from the unstructured litigation data, we used the knowledge model built in step 1 as our structural ontology. Through manual questionnaire entry and crawler codes, we structuralized all the litigation data.

Step 3: Knowledge Fusion

This step solved the problem of inconsistent data quality and structure. We adopted a top-down KG construction method. We used a single data source to avoid, to some extent, such problems as uneven information quality and lack of a hierarchical structure. However, during step 2 (especially with manual data entry), there were differences in understanding among data entry operators. To address these problems, we conducted group training before data entry and answered any questions promptly during the process of data entry. After completing the entry, we undertook data verification to ensure reliability; 20% (1322/6610) of the records were double entered (details are provided in the Graph Theoretical Approaches to Network Analysis section).

Parallel Step: Expert Opinions and Validation

Expert judgment techniques are useful for various reasons, including cost and lack of sufficient observations for quantification with real observed data. We sought expert opinion with the aforementioned 3 steps—especially where little or no data were available for a node or relationship of interest or the existing data were unreliable.

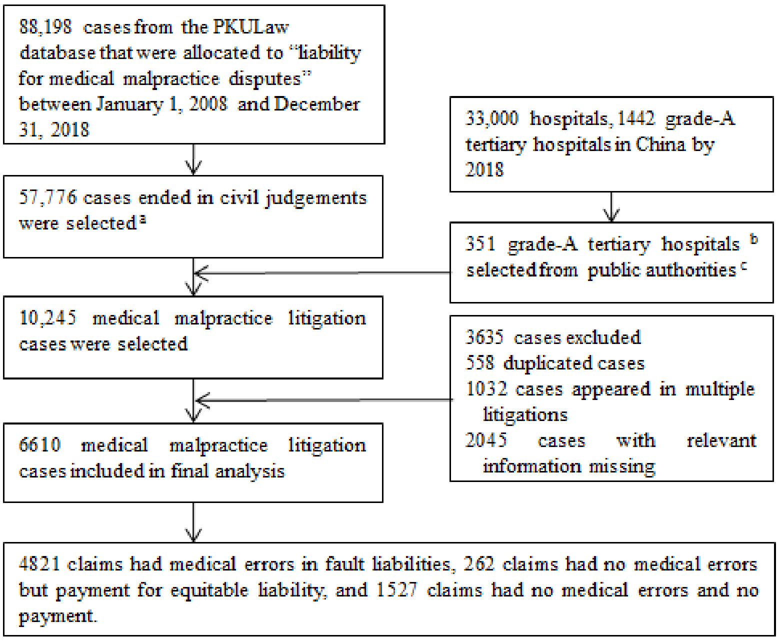
We selected the experts based on their recognized proficiency and experience in medical malpractice, patient safety, KGs, and IT related to this study. We chose our panel of experts from a number of reputable Chinese medical institutions, including the China Hospital Development Institution of Shanghai Jiao Tong University and the School of Public Health of Shanghai Jiao Tong University. All the experts had access to the medical litigation data stored in the PKULaw database and were involved in all stages of modeling, extraction, and fusion.

Data Collection and Preparation

After finalizing the structure of the KG (knowledge model and its graphical representation), we used the available data to quantify the KG. We used the PKULaw database (a publicly available database) as the basis for our study. The database is a national repository of all medical malpractice litigation cases against hospitals and has been admitted by the Supreme Court of China since 2003. As of December 30, 2019, the database covered >76 million litigation cases. All the medical malpractice litigation cases in the database were in text format; however, they all had similar content and structure. Specifically, each case was required to have recorded all the following information: the plaintiff and defendant, any medication involved, any hospital-acquired injury, adverse outcomes, evidence of potential negligence, legal questions, and relevant legislation and judgment.

We searched the PKULaw database and downloaded files on litigation cases that were concluded from January 1, 2008, to December 31, 2018, in the category of “liability for medical malpractice disputes.” The inclusion criteria were (1) cases concluded with a civil judgment and related to grade-A tertiary hospitals and (2) tertiary hospitals on one of the ranking lists published by the Chinese public authorities. We filtered the records using each eligible hospital’s name as a keyword. We excluded records where basic information was missing or duplicate records of individual cases. If a case was reported in multiple records, we kept only the record of the final judgment (Figure 2).

Figure 2. Flowchart of selection of medical malpractice claims in China from 2008 to 2018. a: Civil ligations have three results in China: civil ruling, civil judgment, and civil mediation in court. Cases that end in a civil ruling or mediation do not record relevant information in detail, especially medical information; thus, we excluded such cases. b: Grade-A tertiary hospitals are the highest-level institutions in China. Our selected 351 grade-A tertiary institutions amounted to only 1.1% of all hospitals in China; however, their total number of admissions in 2018 was estimated to be 28 million. We gathered information mainly from the hospitals' official websites. These 28 million admissions accounted for 11% of the nation's total number of admissions in 2018 (254 million, gathered from the China Health Statistics Yearbook [National Health Commission of the People's Republic of China 2019]). c: Eligibility required that a hospital be on a list of public authorities in any previous year. We included four influential ranking lists by public authorities in China: the Best Hospital Ranking by the Hospital Management Institute of Fudan University, the Science and Technology Evaluation Metrics of Hospitals by the Chinese Academy of Medical Sciences, the Hospital Competitiveness Ranking by the Alibi Hospital Management Research Center, and the Best Clinical Specialty Assessment Ranking by Peking University.


88,198 cases from the PKULaw database that were allocated to "liability for medical malpractice disputes" between January 1, 2008 and December 31, 2018
57,776 cases ended in civil judgements were selected a
10,245 medical malpractice litigation cases were selected
6610 medical malpractice litigation cases included in final analysis
33,000 hospitals, 1442 grade-A tertiary hospitals in China by 2018
351 grade-A tertiary hospitals b selected from public authorities c
3635 cases excluded
558 duplicated cases
1032 cases appeared in multiple litigations
2045 cases with relevant information missing
4821 claims had medical errors in fault liabilities, 262 claims had no medical errors but payment for equitable liability, and 1527 claims had no medical errors and no payment.

Graph Theoretical Approaches to Network Analysis

Overview

To investigate networks systematically, we had to define precisely what we meant by "network." In the simplest terms, a network is a collection of *nodes* connected by *relationships*. Nodes correspond to the entities in a network and links to the connections among them [46]. If a network has a large number of nodes with complex relationships, it can be called a *complex network*. In network science, the number of relationships coming into (or out of) a node is called the *degree* of that node—that is the most fundamental network measure; most other measures are ultimately linked to node degree [46].

We examined the network structure to gain greater insight into what we were dealing with. Two types of models are often examined: *random* and *scale-free* networks. Random networks assume that all connections are equally probable, resulting in a Poisson or bell-shaped degree distribution [47]. A scale-free network assumes that the degree distribution follows a power law [35]. In this study, we plotted the *degree distribution* [36] of the MMNC to gain a preliminary understanding of its architecture. We then conducted a scale-free network test, which allowed us to determine the best-fitting power-law model, test its statistical plausibility, and compare it with alternative distributions using a likelihood ratio test [48]. We analyzed the data using R code posted on the web by Clauset et al [48].

We further examined the topological properties of complex systems, such as centrality [49] and distribution of network hubs [50]. The term "hubs" refers to nodes with high degree or high centrality; the removal of hubs can offer advantages with respect to the MMNC. The centrality metrics used in this study included in-degree, closeness, betweenness, and PageRank; they represented a node's distance advantage through its direct connection to others, a node being accessible to others, a node being an intermediary between others, and a node's importance, respectively. In this study, we used the centrality algorithms provided in Neo4j.

Degree Centrality

Degree centrality measures the number of incoming and outgoing relationships of a node and, thus, can help us find popular nodes in a network [35]. The degree centrality of a node *i* reflects its connectivity in the network and is written as $D(i)=d_i/(N-1)$, where *N* is defined as the number of nodes and d_i is defined as the degree of node *i*, that is, the number of incoming and outgoing relationships of node *i*.

Closeness Centrality

Closeness centrality is a way of detecting nodes that are able to spread information very efficiently through a given network. Nodes with a high closeness score have the shortest distances to all other nodes [51], meaning that they are convenient to reach other nodes. The closeness of node *i* is defined as $C(i)=(N-1)/$, where N is defined as the number of nodes and D_{ij} is defined as the shortest path between nodes *i* and *j*. When no path exists between nodes *i* and *j*, D_{ij} is equal to 0.

Betweenness Centrality

Betweenness centrality is a way of detecting the amount of influence a node has on the flow of information in a network, first described by Anthonisse [52] and Freeman [53]. It is often used to find nodes that serve as a bridge from one part of a network to another. For example, people with high betweenness centrality tend to be brokers on social networks by combining different perspectives, transferring ideas between groups. The betweenness of a node *i* reflects its transitivity and is defined as B(i)=, where g_{ab} is the sum of all the shortest paths between nodes *a* and *b*, is the number of the shortest paths that pass through node *i*, and *a*≠*b*≠*i*.

PageRank Centrality

PageRank centrality measures the transitive influence or connectivity of nodes, and it is used to rank websites in Google search results. For example, the home page usually has the highest PageRank centrality as it has incoming links from all other pages. The PageRank score of node *i* counts the number and quality of links to a page, which determines an estimation of how important the page is and is written as PR(i)=(1–*d*)+*d* (PR[T1]/C[T1]+...+PR[Tn]/C[Tn]), where we assume that a page *i* has pages T1 to Tn that point to it and *d* is a damping factor that can be set between 0 and 1. It was set to 0.85 in this study. C(i) is defined as the number of links going out of page *i*.

Ethics Approval

The data used in this study were publicly available and considered “not regulated” by the institutional review boards of the relevant hospitals.

Results

Conceptual Structure of the KG

We abstracted and integrated 8 key concepts and 9 types of relationships into the conceptual graph representation of the MMNC (the overall graph in Figure 3). Multiple medical errors in a case were connected sequentially by the order of occurrence (error subgraph in Figure 3). For instance, patient A had breast cancer, and she also had diabetes. She sought medical services from the oncology department at hospital B. Owing to a delay in treatment and other risk factors, patient A unfortunately died. A malpractice claim was filed, and hospital B paid compensation according to the legal judgment. All the key concepts in the MMNC are defined in Table 1.

The distribution of the number of relationships per node was highly skewed, with a median of 1 relationship per node. The top 0.78% (149/19,099) nodes accounted for most (28,850/57,700, 50%) relationships in the graph. In the graph, 34.45% (6580/19,099) of nodes had only a single relationship.

Figure 3. Conceptual knowledge graph representation of the medical malpractice network of China.

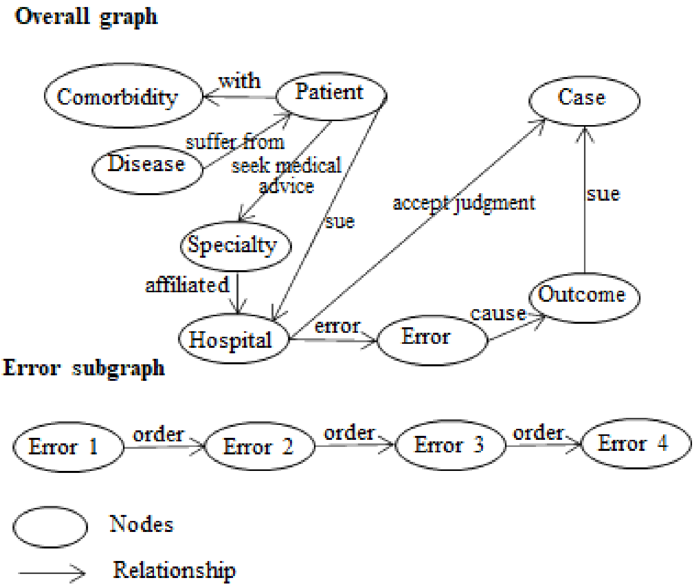


Table 1. Definitions of nodes and relationships.

Name	Type	Definition	Number
P[a]	Node	• Plaintiffs, claims of negligence in the medical service they received • This type of node recorded selected attributes of a patient such as patient demographics.	6582
H[b]	Node	• Defendants offering medical services for plaintiffs • This type of node recorded selected attributes of a hospital such as hospital level or geographic location.	351
S[c]	Node	• Physicians' specialty • This type of node recorded selected attributes of a specialty such as type.	38
O[d]	Node	• The impact on a patient, which is wholly or partially attributable to an error or a series of errors • This type of node recorded the degree of an outcome, which was adapted from Patient Outcome in the WHO-ICPS[e], including the following[f]: • None: patient outcome is not symptomatic, or no symptoms are detected and no treatment is required. • Minor injury: patient outcome is symptomatic, symptoms are mild, loss of function or harm is minimal or intermediate but short term, and no or minimal intervention is required. • Severe injury: patient outcome is symptomatic, requiring life-saving intervention or major surgical or medical intervention, shortening life expectancy, or causing major permanent or long-term harm or loss of function. • Death: on balance of probabilities, death was caused or brought forward in the short term by the error(s). • Mental injury only: patient outcome is only mentally symptomatic, and no other symptoms are detected.	5
C[g]	Node	• Malpractice claims because of professional misconduct or error or demonstration of an unreasonable lack of skill with the result of injury, loss, or damage to the patient • This type of node recorded selected attributes of a claim such as case details or the court.	6610
CD[h]	Node	• Comorbidities according to the CCI[i] • This type of node recorded scores on the CCI.	20
E[j]	Node	• A failure to carry out a planned action as intended or application of an incorrect plan • This type of node recorded types of errors, which was adapted from incident type in the WHO-ICPS and revised by expert opinions, generally classified into "technical error" (related to diagnosis or drugs used) and "nontechnical error" (related to medical records, informed consent, or privacy). More details are provided in Multimedia Appendix 1.	125
D[k]	Node	• Disease groups; diseases were classified into 23 categories according to the ICD-10[l] used by the WHO[m]. • This type of node recorded selected attributes of a disease such as its status and group.	5368
With	Relationship	• Patients' comorbidities; links between P and CD	2097
Suffer from	Relationship	• Patients' disease groups; links between P and D	6610
Seek medical advice	Relationship	• Patients' admission specialties; links between P and S	6610
Affiliated	Relationship	• The subordinate relationship between admission specialties and hospitals; links between H and S	6610
Error	Relationship	• The occurrence of medical errors based on court judgments; links between H and E	4821
Accept judgment	Relationship	• Court decision of malpractice claims; links between H and C	6610
Cause	Relationship	• Hospitals' negligence causes patients' bad outcome; links between O and C	4821
Sue	Relationship	• Patients (with bad outcome) bring hospitals to court; links between O and C or P and H	13,320
Order	Relationship	• The occurrence order of errors; links between E and E	6201

[a] P: patient.

[b] H: hospital.

[c] S: specialty.

[d] O: outcome.

[e] WHO-ICPS: International Classification of Patient Safety from the World Health Organization.

[f] In practice, we measured “Outcome” by combining the 10 types of relationships based on disability levels, which were classified by the Medical Accident Grading Standard in China (for Trial Implementation since 2002), into four categories: minor injury (injury below the disability level of 5), serious injury (disability level of 1 to 5), death, and mental injury only. The more serious the injury, the lower the disability level; disability level 1 is the most serious injury excepting death.

[g] C: case.

[h] CD: comorbidity.

[i] CCI: Charlson Comorbidity Index.

[j] E: error.

[k] D: disease.

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

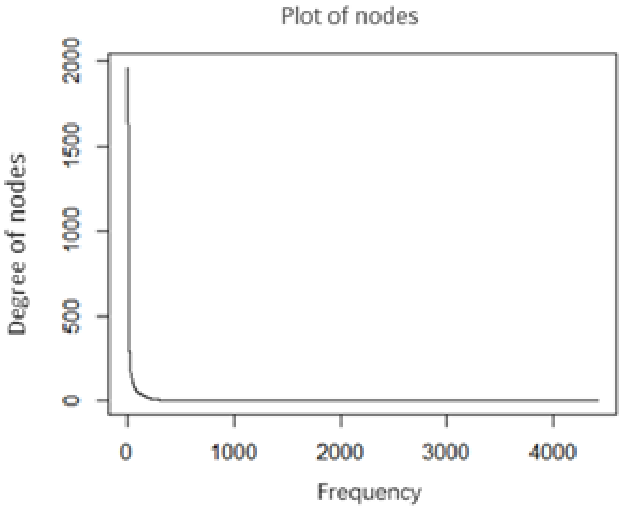
[m] WHO: World Health Organization.

Distribution of the Malpractice Network

In medical malpractice, random events do not occur. The steep curve in Figure 4 shows that the network had many nodes with only a small number of relationships; a few hubs exhibited an extraordinarily large number of relationships. The distinguishing feature of a power law is that there are many small events, and numerous tiny events coexist with a few very large ones. These extraordinarily large events simply do not exist in a bell curve.

In accordance with the method by Clauset [48], we obtained our best-fitting power-law distribution model with the parameters X_{min}=137 and α=2.463458. After we performed 2500 Kolmogorov-Smirnov tests, 2489 (99.56%) failed to reject the scale-free hypothesis. We also fitted an exponential and log-normal distribution to medical malpractice data and performed a goodness-of-fit test to see if these fits were any good. We obtained our best-fitting exponential distribution model with the parameter λ=0.1889905 and our best-fitting log-normal distribution model with the parameters μ=0.5699136 and σ=1.846312. After we performed 2500 Kolmogorov-Smirnov tests for each distribution model, the results were similar; that is, 100% (2500/2500) rejected the scale-free hypothesis. We concluded that the power-law distribution displayed a good fit to the degree distribution of nodes from the MMNC (ie, it was a scale-free network).

Figure 4. Degree distribution of the network.



Hub Nodes in the Malpractice Network

Scale-free networks are characterized by high clustering and skewed degree distributions. Such features predict that each scale-free network will have several large hubs that will fundamentally define the network’s topology (Tables 2 and 3). More information about the sample characteristics is provided in Multimedia Appendix 2.

Table 2 reports the top 10 nodes by degree, closeness, betweenness, and PageRank. Orthopedics, obstetrics and gynecology, emergency medicine, gastroenterology, general surgery, and cancer were ranked as the top specialties in all 4 metrics. On the basis of degree, betweenness, and PageRank, the 3 outcome nodes for death, minor injury, and severe injury were ranked close to the forefront. Specific medical errors appear a number of times in Table 2: inadequate informed consent, delay in treatment, and failure to recognize complications. In general, the results of the 4 centrality metrics were relatively consistent; the nodes that were ranked at the top had a higher degree of coincidence.

Similarly, Table 3 indicates that a few nontechnical errors such as inadequate informed consent and illegible medical records appeared as the top errors with almost all metrics. However, in terms of betweenness, technical errors (including delay in treatment and failure to recognize complications) had higher values. All the top 10 errors with the PageRank metric were nontechnical. In general, the error nodes that were ranked high were relatively consistent, and nontechnical errors were more central than technical errors.

Table 2. Top 10 nodes by degree, closeness, betweenness, and PageRank in the overall graph.[a]

Rank	Degree	Closeness	Betweenness	PageRank
1	Death	Orthopedics	Death	Death
2	Minor injury	Emergency medicine	Orthopedics	Minor injury
3	Severe injury	Failure to perform preoperative evaluation[b]	Minor injury	Severe injury
4	Orthopedics	Missed diagnosis[b]	Emergency medicine	Orthopedics
5	Inadequate informed consent[b]	Obstetrics and gynecology	Obstetrics and gynecology	Inadequate informed consent[b]
6	Obstetrics and gynecology	Delay in diagnosis[b]	Gastroenterology	Obstetrics and gynecology
7	Emergency medicine	Gastroenterology	Inadequate informed consent[b]	Delay in treatment[b]
8	Other comorbidities	Inadequate informed consent[b]	Severe injury	Emergency medicine
9	Gastroenterology	Failure to recognize complications[b]	Cancer	Gastroenterology
10	Delay in treatment[b]	Cancer	General surgery	Failure to recognize complications[b]

[a]The definitions of all the nodes can be found in Table 1.

[b]These are error nodes; all errors are described in Multimedia Appendix 1.

Table 3. Top 10 errors by degree, closeness, betweenness, and PageRank in the error subgraph.[a]

Rank	Degree	Closeness	Betweenness	PageRank
1	Inadequate informed consent	Inadequate informed consent	Delay in treatment[b]	Inadequate informed consent
2	Unclear, ambiguous, illegible, or incomplete medical records	Unclear, ambiguous, illegible, or incomplete medical records	Lack of informed consent	Supervision or patient safety management
3	Supervision or patient safety management	Delay in treatment[b]	Failure to recognize complications[b]	Unclear, ambiguous, illegible, or incomplete medical records
4	Delay in treatment[b]	Failure to perform preoperative evaluation[b]	Failure to perform preoperative evaluation[b]	Failure to communicate with or instruct the patient or family
5	Failure to recognize complications[b]	Supervision or patient safety management	Unclear, ambiguous, illegible, or incomplete medical records	Lack of informed consent
6	Lack of informed consent	Failure to perform pretreatment evaluation[b]	Untimely patient rounds[b]	Emergency management
7	Failure to communicate with or instruct the patient or family	Failure to identify postoperative complications[b]	Failure to perform pretreatment evaluation[b]	Unsigned consent documentation
8	Failure to perform pretreatment evaluation[b]	Lack of informed consent	Delay in diagnosis[b]	Administrative management
9	Other surgery-related errors[b]	Delay in diagnosis[b]	Delay in surgery[b]	Other management-related errors
10	Other treatment-related errors[b]	Delay in surgery[b]	Other medicine-related errors[b]	Risk management

[a]All errors are described in Multimedia Appendix 1.

[b]Attributed to technical errors.

Discussion

Principal Findings

This study constructed a KG derived from medical malpractice litigation data to represent the MMNC. We found that the MMNC was a scale-free network instead of a random network. Scale-free networks representing the MMNC were high clustering, showed skewed degree distributions, and had hub nodes. The results of the hub nodes revealed that orthopedics, obstetrics and gynecology, and the emergency department were the 3 most frequent specialties that incurred medical malpractice; inadequate informed consent work constituted the most errors. Nontechnical errors (eg, inadequate informed consent) showed a higher centrality than technical errors.

Power laws are being discovered in a great number and with various phenomena; accordingly, some authors have described them as “more normal than ‘normal’” [54]. Power laws rarely emerge in systems completely dominated by a roll of the dice [55]. Thus, the power law that we observed with the MMNC signified that real networks are far from random. Plausible explanations for the nonrandom nature of the MMNC described in this study include the involvement of various human factors or errors. In the United States, the National Practitioner Data Bank classifies medical errors according to malpractice allegations, but subclassified terms are not further defined [56]. Numerous studies [57,58] have investigated the causal factors of medical malpractice by developing various human factor classification frameworks. Many countries have established adverse event reporting systems and classified those events—it is on such classification that the WHO-ICPS, referenced in this study, is based. However, those classification frameworks have not been widely used worldwide, and some frameworks have yet to be improved.

In complex theory, the widely accepted explanation for the existence of most (if not all) scale-free networks in the real world is growth and preferential attachment (ie, a particular growth process for such networks), as proposed by Barabási and Albert [35]. Thus, each network starts with a core node and grows by adding new nodes. There are connections among nodes—as more nodes become connected, the number of connections that result is greater. In the context of medical malpractice, the more hospitals or physicians with poor malpractice records, the greater the likelihood that they will become involved in future such cases. This is in harmony with the idea of the Pareto law or principle, which is also known as the 80/20 rule [59]. Accordingly, there has to be some order behind these complex systems [46,55]. The causes of the power law found in the MMNC need to be further studied.

The network analysis help identify hub nodes for interventions. The inevitability of the existence of hub nodes in scale-free networks presents an opportunity for prevention and control of medical malpractice. Consistent with the findings of recent research [2-4,13,14,19,20,23], we found that specialties such as orthopedics, obstetrics and gynecology, and the emergency department incur a disproportionately large share of litigation cases. The specific reasons are unknown; however, potential explanations are that such specialties admit higher-risk patients, operate in higher-risk environments, or are subject to the “bad apple effect” (ie, repeatedly provide substandard care) [60]. The hospitals included in this study are the top tertiary hospitals across China compared with other levels of medical institutions, which have better medical resources and treat more patients with intractable diseases. Some specific specialties of these hospitals are more likely to have a high incidence of medical malpractice. Obstetrics and gynecology involves the health of both newborns and *puerpera*, whereas orthopedic diseases have a more intuitive impact on limb function and daily work. Patients with orthopedic diseases tend to expect dramatic improvements in limb function following a major procedure, but unsatisfactory treatment results might occur. Emergency patients tend to have acute onset or severe illness, especially when there is no family member around to sign the informed consent, and the risk of medical malpractice in such cases could be higher. The “bad apple effect” could be explained by the anchoring effect; that is, because of the cognitive errors, medical staff might repeatedly provide substandard care with certain medical errors. The cognitive errors might have formed from previously acquired information or experience, and such errors are like an anchor sinking to the bottom of the sea, holding medical staff’s thoughts in place. In fact, it is what we often refer to as a “preconceived” notion.

Compared with technical errors, nontechnical errors had greater centrality in this study. However, descriptive studies in this field [13,14] show that technical errors occur more frequently. Our findings suggest that it may be effective to improve nontechnical skills to reduce accidents [61]. Our findings demonstrated that one of the most prominent nontechnical errors involved inadequate informed consent. Informed consent has always been one of the most common medical errors in China. In total, 2 Chinese studies [62,63] found that 23% to 43% of medical lawsuits involved incomplete consent notification for patients. Owing to the information asymmetry between physicians and patients, coupled with the tense relationship between physicians and patients in China [18], patients’ doubts will trigger medical malpractice once medical staff are insufficient in risk notification. In addition, errors related to medical records were particularly prominent among nontechnical errors. A plausible explanation is that medical records are the main evidence in the mediation of medical malpractice in China, and irregular writing will directly affect the judgment of medical litigation [14].

We found that the dominant factors in technical errors were inadequate attention and delays, including treatment delay, failure to recognize complications, and delays in surgery and diagnosis. Unlike in the United States, where diagnostic errors are the most common cause of malpractice claims [64,65], treatment and surgical errors are more frequent in China. The difference may be due to the fact that the medical system in the United States may be relatively fragmented (eg, the diagnosis and treatment of the same patient may be divided into different institutions, resulting in medical staff often diagnosing based on more fragmented information). Diagnostic errors may be ignored in China as medication and surgical errors are more easily observed during medical treatment. There is still considerable room in China for enhancing the quality of health care and patient safety management. There are variations in trends of technical errors in different specialties in China; however, there may be common interventions for nontechnical errors. For example, shared decision-making approaches can be and have been applied to all specialties; this helps protect physicians from malpractice claims and ensures that patients are better informed [66].

This study has found a number of hub nodes in the MMNC, including technical and nontechnical errors, which could be helpful for preventive education for medical malpractice. Nontechnical errors occupy an important position in the MMNC, reflecting the lack of awareness of error prevention in medical institutions and their medical staff. Compared with technical errors, nontechnical errors related to informed consent notification, physician-patient communication skills, and medical record writing could be relatively easily avoided by strengthening related training. However, the education of medical students in China places the most emphasis on clinical skills and scientific research, and training to avoid medical errors, especially nontechnical errors, is very limited. We believe that medical education and training should be strengthened to constantly improve clinical performance and the awareness of nontechnical errors among medical students and staff.

Network analysis provides a useful tool for analyzing medical malpractice. It does not require a complete map of medical malpractice, only measuring the degree distribution by analyzing a representative subset of the complete network [55]; we do so in this study. It is impossible to obtain medical malpractice data without omissions and build a complete malpractice network. Fortunately, a complete map of medical malpractice is not necessary to determine whether it is scale-free or random [55]. Another problem is identifying the hubs—doubtlessly, many hubs may have gone undiscovered in this study, and we may have included a few nonhubs. Decades of research have produced numerous graph methods for identifying hubs. Such methods may be imperfect, but they are still useful—it is possible to identify the hubs with a certain probability. Dezső and Barabási [10] demonstrated that any policy that displayed bias toward more connected nodes—even a small bias—restored the finite epidemic threshold. In the context of malpractice, it may not be possible to find all the hubs; however, by attempting to do so, the spread of medical malpractice can be limited. Network analysis is an emerging research field that has grown with the development of network theory and computer technology. In the real world, there are many fields that can be abstracted into complex networks. Physicists have found that power laws frequently signal a transition from disorder to order—such a distribution pattern is observed in most self-organized complex systems in nature, technology, and society [46,55]. Many people feel that they do not live in a random world—there have to be certain key organizational principles behind complex systems. Finding the rules hidden behind the structure in the MMNC is the next future direction.

Limitations

This study had several limitations. First, medical malpractice litigation cases presumably represent the tip of the iceberg with medical errors, in which patients receive poor-quality health care [67]. Second, we assumed that the Chinese judiciary system is fair, independent, and strong; however, there are several deficiencies or flaws in medical malpractice law in China. Finally, simplified network models cannot explain everything regarding their real-world counterparts. With the MMNC, we assumed that all the nodes were identical except for their degree and that all links were of the same type and had the same strength; however, that is not the case in real-world networks.

Conclusions

This study constructed a KG derived from medical malpractice litigation data to represent the MMNC. We demonstrated that it was a scale-free network, not a random network, and showed that the occurrence of medical malpractice was traceable. The MMNC was in transition from chaos to order, reflecting from the results of the hub nodes that there were several key specialties and errors. Faced with limited resources, it is necessary to make specific interventions for key specialties and errors as well as pay greater attention to nontechnical errors; doing so could effectively control medical risks.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China for their financial support. They also appreciate the 13th International Conference on Health Policy Statistics, which offered an opportunity for a poster presentation of this paper. This study was supported by the National Natural Science Foundation of China (grant 72074147). The funding agency did not play a role in designing and conducting this research or analyzing the data. The decision to submit the manuscript for publication was that of the authors.

Data Availability

Deidentified individual participant data that underlie the results reported in this paper (tables, figures, and multimedia appendices) can be made available to researchers who apply for proposed use after approval by the corresponding author of this study. A signed data access agreement is also needed for data requesters.

Authors' Contributions

CS and GL were responsible for the design of the study. Data collection and analysis were initially carried out by SD under the guidance of CS. The paper was primarily written by SD. All the authors commented on successive drafts of the paper and suggested areas for revision and improvement. All the authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of all medical errors.
[DOCX File , 28 KB - medinform_v10i11e35709_app1.docx]

Multimedia Appendix 2
Distribution of malpractice claims characteristics by data sample.
[DOCX File , 17 KB - medinform_v10i11e35709_app2.docx]

References

1. Studdert D, Spittal M, Zhang Y, Wilkinson DS, Singh H, Mello MM. Changes in practice among physicians with malpractice claims. N Engl J Med 2019 Mar 28;380(13):1247-1255 [FREE Full text] [doi: 10.1056/NEJMsa1809981] [Medline: 30917259]
2. Carrier E, Reschovsky J, Katz D, Mello MM. High physician concern about malpractice risk predicts more aggressive diagnostic testing in office-based practice. Health Aff (Millwood) 2013 Aug;32(8):1383-1391 [FREE Full text] [doi: 10.1377/hlthaff.2013.0233] [Medline: 23918482]
3. Studdert D, Bismark M, Mello M, Singh H, Spittal MJ. Prevalence and characteristics of physicians prone to malpractice claims. N Engl J Med 2016 Jan 28;374(4):354-362 [FREE Full text] [doi: 10.1056/nejmsa1506137]
4. Schaffer A, Jena A, Seabury S, Singh H, Chalasani V, Kachalia A. Rates and characteristics of paid malpractice claims among US physicians by specialty, 1992-2014. JAMA Intern Med 2017 May 01;177(5):710-718 [FREE Full text] [doi: 10.1001/jamainternmed.2017.0311] [Medline: 28346582]
5. Black B, Hyman D, Lerner J. Physicians with multiple paid medical malpractice claims: are they outliers or just unlucky? Int Rev Law Econ 2019 Jun;58:146-157 [FREE Full text] [doi: 10.1016/j.irle.2019.03.006]
6. Hofmann S, Curtiss J, McNally R. A complex network perspective on clinical science. Perspect Psychol Sci 2016 Sep;11(5):597-605 [FREE Full text] [doi: 10.1177/1745691616639283] [Medline: 27694457]
7. Liljeros F, Edling C, Amaral L, Stanley HE, Aberg Y. The web of human sexual contacts. Nature 2001 Jun 21;411(6840):907-908 [FREE Full text] [doi: 10.1038/35082140] [Medline: 11418846]
8. Schneider J, Zhou A, Laumann E. A new HIV prevention network approach: sociometric peer change agent selection. Soc Sci Med 2015 Jan;125:192-202 [FREE Full text] [doi: 10.1016/j.socscimed.2013.12.034] [Medline: 24518188]
9. Schneeberger A, Mercer CH, Gregson SA, Ferguson NM, Nyamukapa CA, Anderson RM, et al. Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. Sex Transm Dis 2004 Jun;31(6):380-387. [doi: 10.1097/00007435-200406000-00012] [Medline: 15167650]
10. Dezso Z, Barabási AL. Halting viruses in scale-free networks. Phys Rev E Stat Nonlin Soft Matter Phys 2002 May 21;65(5 Pt 2):055103. [doi: 10.1103/PhysRevE.65.055103] [Medline: 12059627]
11. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys Rev Lett 2001 Apr 02;86(14):3200-3203. [doi: 10.1103/PhysRevLett.86.3200] [Medline: 11290142]
12. China judgments online homepage. China Judgments Online. URL: http://wenshu.court.gov.cn/ [accessed 2019-12-29]
13. Wang Z, Li N, Jiang M, Dear K, Hsieh C. Records of medical malpractice litigation: a potential indicator of health-care quality in China. Bull World Health Organ 2017 Mar 13;95(6):430-436 [FREE Full text] [doi: 10.2471/blt.16.179143]
14. Li H, Dong S, Liao Z, Yao Y, Yuan S, Cui Y, et al. Retrospective analysis of medical malpractice claims in tertiary hospitals of China: the view from patient safety. BMJ Open 2020 Sep 24;10(9):e034681 [FREE Full text] [doi: 10.1136/bmjopen-2019-034681] [Medline: 32973050]
15. Pukk-Härenstam K, Ask J, Brommels M, Thor J, Penaloza RV, Gaffney FA. Analysis of 23 364 patient-generated, physician-reviewed malpractice claims from a non-tort, blame-free, national patient insurance system: lessons learned from Sweden. Postgrad Med J 2009 Feb 01;85(1000):69-73. [doi: 10.1136/qshc.2007.022897] [Medline: 19329700]
16. Tilma J, Nørgaard M, Mikkelsen KL, Johnsen SP. Existing data sources for clinical epidemiology: the Danish patient compensation association database. Clin Epidemiol 2015 Jul;2015:7:347-353 [FREE Full text] [doi: 10.2147/clep.s84162]
17. Hesketh T, Wu D, Mao L, Ma N. Violence against doctors in China. BMJ 2012 Sep 07;345(sep07 1):e5730. [doi: 10.1136/bmj.e5730] [Medline: 22960376]
18. Cai R, Tang J, Deng C, Lv G, Pan J. Serious workplace violence against health-care workers in China: synthesising a profile of evidence from national judgment documents. Lancet 2018 Oct;392:S46 [FREE Full text] [doi: 10.1016/s0140-6736(18)32675-8]
19. Hendee WR. To err is human: building a safer health system. J Vascular Intervention Radiol 2001 Jan;12(1):P112-P113 [FREE Full text] [doi: 10.1016/s1051-0443(01)70072-3]
20. Wu C, Weng H, Chen R. Time trends of assessments for medical dispute cases in Taiwan: a 20-year nationwide study. Intern Med J 2013 Sep;43(9):1023-1030 [FREE Full text] [doi: 10.1111/imj.12105] [Medline: 23425553]

21. Jena A, Seabury S, Lakdawalla D, Chandra A. Malpractice risk according to physician specialty. N Engl J Med 2011 Aug 18;365(7):629-636 [FREE Full text] [doi: 10.1056/nejmsa1012370]
22. Hwang C, Wu C, Cheng F, Yen YL, Wu KH. A 12-year analysis of closed medical malpractice claims of the Taiwan civil court: a retrospective study. Medicine (Baltimore) 2018 Mar;97(13):e0237 [FREE Full text] [doi: 10.1097/MD.0000000000010237] [Medline: 29595675]
23. Studdert D, Mello M, Gawande A, Gandhi TK, Kachalia A, Yoon C, et al. Claims, errors, and compensation payments in medical malpractice litigation. N Engl J Med 2006 May 11;354(19):2024-2033 [FREE Full text] [doi: 10.1056/nejmsa054479]
24. Jena A, Schoemaker L, Bhattacharya J, Seabury SA. Physician spending and subsequent risk of malpractice claims: observational study. BMJ 2015 Nov 04;351:h5516 [FREE Full text] [doi: 10.1136/bmj.h5516] [Medline: 26538498]
25. Lopez J, Hollier L. Review of "multisource evaluation of surgeon behavior is associated with malpractice claims" by Lagoo j et al in Ann Surg 270. J Craniofacial Surg 2020;31(3):885-890 [FREE Full text] [doi: 10.1097/scs.0000000000005956]
26. Shouhed D, Beni C, Manguso N, IsHak WW, Gewertz BL. Association of emotional intelligence with malpractice claims: a review. JAMA Surg 2019 Mar 01;154(3):250-256 [FREE Full text] [doi: 10.1001/jamasurg.2018.5065] [Medline: 30698614]
27. Greenberg M, Haviland AM, Ashwood JS, Main R. Is better patient safety associated with less malpractice activity?: evidence from California. Rand Health Q 2011;1(1):1 [FREE Full text] [Medline: 28083157]
28. Minami CA, Chung JW, Holl JL, Bilimoria KY. Impact of medical malpractice environment on surgical quality and outcomes. J Am Coll Surg 2014 Feb;218(2):271-8.e1 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2013.09.007] [Medline: 24211056]
29. Black B, Hyman D, Silver C, Sage W. Defense costs and insurer reserves in medical malpractice and other personal injury cases: evidence from Texas, 1988-2004. Am Law Econ Rev 2008 Aug 01;10(2):185-245 [FREE Full text] [doi: 10.1093/aler/ahn014]
30. Mello M, Chandra A, Gawande AA, Studdert DM. National costs of the medical liability system. Health Aff (Millwood) 2010 Sep;29(9):1569-1577 [FREE Full text] [doi: 10.1377/hlthaff.2009.0807] [Medline: 20820010]
31. Kessler D, McClellan M. Do doctors practice defensive medicine? Q J Econ 1996 May 01;111(2):353-390 [FREE Full text] [doi: 10.2307/2946682]
32. Quinn R. Medical malpractice insurance: the reputation effect and defensive medicine. J Risk Insurance 1998 Sep;65(3):467. [doi: 10.2307/253660]
33. Renkema E, Broekhuis M, Ahaus K. Conditions that influence the impact of malpractice litigation risk on physicians' behavior regarding patient safety. BMC Health Serv Res 2014 Jan 25;14:38 [FREE Full text] [doi: 10.1186/1472-6963-14-38] [Medline: 24460754]
34. Plsek P, Greenhalgh T. Complexity science: the challenge of complexity in health care. BMJ 2001 Sep 15;323(7313):625-628 [FREE Full text] [doi: 10.1136/bmj.323.7313.625] [Medline: 11557716]
35. Barabasi A, Albert R. Emergence of scaling in random networks. Science 1999 Oct 15;286(5439):509-512 [FREE Full text] [doi: 10.1126/science.286.5439.509] [Medline: 10521342]
36. Amaral LA, Scala A, Barthelemy M, Stanley HE. Classes of small-world networks. Proc Natl Acad Sci U S A 2000 Oct 10;97(21):11149-11152 [FREE Full text] [doi: 10.1073/pnas.200327197] [Medline: 11005838]
37. Xiao Y, Zhao K, Bishai D, Peters DH. Essential drugs policy in three rural counties in China: what does a complexity lens add? Soc Sci Med 2013 Sep;93:220-228 [FREE Full text] [doi: 10.1016/j.socscimed.2012.09.034] [Medline: 23103350]
38. Sporns O, Chialvo DR, Kaiser M, Hilgetag CC. Organization, development and function of complex brain networks. Trends Cogn Sci 2004 Sep;8(9):418-425. [doi: 10.1016/j.tics.2004.07.008] [Medline: 15350243]
39. Bassett D, Bullmore E. Small-world brain networks. Neuroscientist 2006 Dec;12(6):512-523 [FREE Full text] [doi: 10.1177/1073858406293182] [Medline: 17079517]
40. Reijneveld J, Ponten S, Berendse H, Stam C. The application of graph theoretical analysis to complex networks in the brain. Clin Neurophysiol 2007 Nov;118(11):2317-2331 [FREE Full text] [doi: 10.1016/j.clinph.2007.08.010] [Medline: 17900977]
41. Stam C, Reijneveld J. Graph theoretical analysis of complex networks in the brain. Nonlinear Biomed Phys 2007 Jul 05;1(1):3 [FREE Full text] [doi: 10.1186/1753-4631-1-3] [Medline: 17908336]
42. Barth lemy M. Betweenness centrality in large complex networks. Eur Physical J B - Condensed Matter 2004 Mar 1;38(2):163-168 [FREE Full text] [doi: 10.1140/epjb/e2004-00111-4]
43. Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci U S A 2002 Jun 11;99(12):7821-7826 [FREE Full text] [doi: 10.1073/pnas.122653799] [Medline: 12060727]
44. Blazing-fast graph, petabyte scale. Neo4j. URL: http://www.neo4j.org/ [accessed 2021-12-26]
45. Runciman W, Baker G, Michel P, Dovey S, Lilford RJ, Jensen N, et al. Tracing the foundations of a conceptual framework for a patient safety ontology. Qual Saf Health Care 2010 Dec;19(6):e56 [FREE Full text] [doi: 10.1136/qshc.2009.035147] [Medline: 20702442]
46. Mitchell M. Complexity A Guided Tour. USA: Oxford University Press; 2009.
47. Bollobás B. Random Graphs Second Edition. Cambridge, UK: Cambridge University Press; 2011.
48. Clauset A, Shalizi C, Newman M. Power-law distributions in empirical data. SIAM Rev 2009 Nov 04;51(4):661-703 [FREE Full text] [doi: 10.1137/070710111]

49. Grando F, Noble D, Lamb LC. An analysis of centrality measures for complex and social networks. In: Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM). 2016 Presented at: 2016 IEEE Global Communications Conference (GLOBECOM); Dec 04-08, 2016; Washington, DC, USA URL: https://doi.org/10.1109/GLOCOM.2016.7841580 [doi: 10.1109/glocom.2016.7841580]
50. Wang J, Mo H, Wang F, Jin F. Exploring the network structure and nodal centrality of China's air transport network: a complex network approach. J Transport Geography 2011 Jul;19(4):712-721 [FREE Full text] [doi: 10.1016/j.jtrangeo.2010.08.012]
51. Sabidussi G. The centrality of a graph. Psychometrika 1966 Dec;31(4):581-603 [FREE Full text] [doi: 10.1007/BF02289527] [Medline: 5232444]
52. Anthonisse J. The Rush in a Directed Graph. Amsterdam, Netherlands: Stichting Mathematisch Centrum. Mathematische Besliskunde; 1971.
53. Freeman L. A set of measures of centrality based on betweenness. Sociometry 1977 Mar;40(1):35 [FREE Full text] [doi: 10.2307/3033543]
54. Willinger W, Alderson D, Doyle J, Li L. More "normal" than normal: scaling distributions and complex systems. In: Proceedings of the 2004 Winter Simulation Conference, 2004. 2004 Presented at: Proceedings of the 2004 Winter Simulation Conference, 2004; Dec 05-08, 2004; Washington, DC, USA URL: https://doi.org/10.1109/WSC.2004.1371310 [doi: 10.1109/WSC.2004.1371310]
55. Mackay R. Linked: how everything is connected to everything else and what it means for business, science and everyday life. Complicity Int J Complexity Educ 2005 Dec 01;2(1). [doi: 10.29173/cmplct8735]
56. National Practitioner Data Bank Public Use Data File. U.S. Department of Health and Human Services. URL: https://www.npdb-hipdb.com/statistical-data/public-use-data-file/ [accessed 2022-10-23]
57. Pinto A, Scuderi M, Daniele S. Errors in radiology: definition and classification. In: Errors in Radiology. Milano: Springer; 2012.
58. Taib IA, McIntosh AS, Caponecchia C, Baysari M. A review of medical error taxonomies: a human factors perspective. Safety Sci 2011 Jun;49(5):607-615 [FREE Full text] [doi: 10.1016/j.ssci.2010.12.014]
59. Newman M. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 2005 Sep;46(5):323-351 [FREE Full text] [doi: 10.1080/00107510500052444]
60. Sloan FA, Mergenhagen PM, Burfield WB, Bovbjerg RR, Hassan M. Medical malpractice experience of physicians. Predictable or haphazard? JAMA 1989 Dec 15;262(23):3291-3297. [Medline: 2585673]
61. Uramatsu M, Fujisawa Y, Mizuno S, Souma T, Komatsubara A, Miki T. Do failures in non-technical skills contribute to fatal medical accidents in Japan? A review of the 2010-2013 national accident reports. BMJ Open 2017 Feb 16;7(2):e013678 [FREE Full text] [doi: 10.1136/bmjopen-2016-013678] [Medline: 28209605]
62. Big data report on national medical malpractice cases in 2019. CN-Healthcare. URL: https://www.cn-healthcare.com/articlewm/20200224/content-1090334.html [accessed 2022-10-23]
63. Big data report on national medical damage liability dispute cases in 2018. Sohu. URL: https://www.sohu.com/a/292549938_258430 [accessed 2020-02-24]
64. Saber Tehrani AS, Lee H, Mathews SC, Shore A, Makary MA, Pronovost PJ, et al. 25-Year summary of US malpractice claims for diagnostic errors 1986-2010: an analysis from the National Practitioner Data Bank. BMJ Qual Saf 2013 Aug;22(8):672-680. [doi: 10.1136/bmjqs-2012-001550] [Medline: 23610443]
65. Gupta A, Snyder A, Kachalia A, Flanders S, Saint S, Chopra V. Malpractice claims related to diagnostic errors in the hospital. BMJ Qual Saf 2017 Aug 09;27(1):53-60. [doi: 10.1136/bmjqs-2017-006774] [Medline: 28794243]
66. Birkeland SF. Informed consent obtainment, malpractice litigation, and the potential role of shared decision-making approaches. Eur J Health Law 2017 May 31;24(3):264-284. [doi: 10.1163/15718093-12341410]
67. Garon-Sayegh P. Analysis of medical malpractice claims to improve quality of care: cautionary remarks. J Eval Clin Pract 2019 Oct 09;25(5):744-750. [doi: 10.1111/jep.13178] [Medline: 31069900]

Abbreviations

KG: knowledge graph
MMNC: medical malpractice network of China
WHO-ICPS: International Classification of Patient Safety from the World Health Organization

Edited by C Lovis; submitted 14.12.21; peer-reviewed by W Ceron, G Mayer; comments to author 10.07.22; revised version received 28.08.22; accepted 24.09.22; published 03.11.22.

Please cite as:
Dong S, Shi C, Zeng W, Jia Z, Dong M, Xiao Y, Li G
The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study
JMIR Med Inform 2022;10(11):e35709
URL: https://medinform.jmir.org/2022/11/e35709
doi:10.2196/35709
PMID:36326815

Original Paper

Managing Critical Patient-Reported Outcome Measures in Oncology Settings: System Development and Retrospective Study

Olga Strachna[1,2], MSc; Onur Asan[1], PhD; Peter D Stetson[2], MD, MA

[1]School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ, United States

[2]Division of Digital Products and Informatics, Memorial Sloan Kettering Cancer Center, New York, NY, United States

Corresponding Author:
Onur Asan, PhD
School of Systems and Enterprises
Stevens Institute of Technology
1 Castle Terrace
Hoboken, NJ, 07030
United States
Phone: 1 201 216 5514
Email: oasan@stevens.edu

Abstract

Background: Remote monitoring programs based on the collection of patient-reported outcome (PRO) data are being increasingly adopted in oncology practices. Although PROs are a great source of patient data, the management of critical PRO data is not discussed in detail in the literature.

Objective: This first-of-its-kind study aimed to design, describe, and evaluate a closed-loop alerting and communication system focused on managing PRO-related alerts in cancer care.

Methods: We designed and developed a novel solution using an agile software development methodology by incrementally building new capabilities. We evaluated these new features using participatory design and the Fit between Individuals, Task, and Technology framework.

Results: A total of 8 questionnaires were implemented using alerting features, resulting in an alert rate of 7.82% (36,838/470,841) with 13.28% (10,965/82,544) of the patients triggering at least one alert. Alerts were reviewed by 501 staff members spanning across 191 care teams. All the alerts were reviewed with a median response time of 1 hour (SD 185 hours) during standard business hours. The most severe (red) alerts were documented 56.83% (2592/4561) of the time, whereas unlabeled alerts were documented 27.68% (1298/4689) of the time, signaling clinician concordance with the alert thresholds.

Conclusions: A PRO-based alert and communication system has some initial benefits in reviewing clinically meaningful PRO data in a reasonable amount of time. We have discussed key system design considerations, workflow integration, and the mitigation of potential impact on the burden of care teams. The introduction of a PRO-based alert and communication system provides a reliable mechanism for care teams to review and respond to patient symptoms quickly. The system was standardized across many different oncology settings, demonstrating system flexibility. Future studies should focus on formally evaluating system usability through qualitative methods.

(JMIR Med Inform 2022;10(11):e38483) doi:10.2196/38483

KEYWORDS

patient-reported outcome measures; delivery of health care; self-report; quality of care; neoplasms; surveys and questionnaires; clinical decision support systems

Introduction

Background

Patient-reported outcomes (PROs) are being increasingly collected as a part of routine clinical care, capturing patients' self-reported symptoms, function, and quality of life. They support the goal of facilitating clinician-patient communication, mutual understanding of patient preferences, and enabling shared decision-making with an impact on treatment decisions [1-7]. PRO data collection is particularly significant for the oncology patient population, especially for patients on clinical trials, because of the critical need to track symptomatic adverse events related to cancer treatments, which have a significant impact on the clinical outcomes and quality of life of patients [5,8-14]. In addition, PROs are relied on for managing health care use [4,8,11]. PRO data are only valuable as long as patients complete the surveys accurately and timely. Limited survey completion rates minimize the ability to draw clinical conclusions from sparsely filled out data [15]. Studies have shown that clinician engagement in the process of administering PRO programs via patient education or outreach has a direct positive impact on patient engagement [16-18]. Provider disengagement in the process of reviewing the data may disincentivize patients from completing their PRO assessments [19]. Therefore, it is essential for clinicians to follow-up with patients regarding any significant outcomes reported in a streamlined and timely manner.

As with many other clinical applications, there has been an interest to integrate PROs within clinical workflows; however, evidence shows limited success [14,20-24]. Sources of patient-generated health data such as PROs are relatively new to the standard of care practices, and there is not always a standard mechanism in place for clinicians to handle PROs appropriately with varying implementation strategies [9]. At the same time, notifying clinicians about all patient responses does not always result in a timely follow-up with the patient if there are workflow barriers impeding communication, such as reviewing too many PRO responses [25-27].

The concept of alerting is not new to health care, with long-standing applications in clinical decision support systems for drug-drug interactions [28,29], adverse event monitoring [20,30,31], abnormal laboratory results [25,32-36], and many others [37]. The idea of PRO alerts is distinct from the standard clinical alerts mentioned, in that it involves asynchronous interruptive and noninterruptive communication between patients and care team members as well as coordination among care team members within the system. Several studies have mentioned using alert-based features within the context of PROs; however, none of them have discussed the communication aspect with patients, analyzed the impact on workload, and described the detailed designs of such alerting systems [8,9,38-40].

Given the rise in the popularity of remote monitoring programs, including the use of PRO data during the COVID-19 pandemic, there have been several enthusiastic studies on program evaluations, and remote monitoring programs are expected to increase in adoption in the post–COVID-19 pandemic years [41-44]. Remote monitoring programs are novel in and of themselves; therefore, as a part of this study, it was important to consider the design aspect of a work management system to handle critical results in a timely manner. In addition, it was critical to understand the impact of running such programs asynchronously from clinical visits to allocate appropriate resources to respond to patient-specific needs outside standard staffed business hours, with implications for program monitoring and management.

Objectives

To date, there has been pervasive interest in using PROs for remote symptom management in oncology standard of care practice, but very little is known about the management of critical patient symptom responses and the engagement of clinical staff in the review of patient responses to address them appropriately. Given the increasing adoption of PROs in standard oncological practices, we identified a need to design a robust PRO alert management and communication system that scales with increasing clinic patient volumes and patient demand for asynchronous communication. Considering the potential clinic disruption, it was important to quantify the impact of such a new system on clinic workloads. For this study, we designed and implemented the alert and communication system separate from the electronic health record (EHR) but with a tight integration of key results. It was unknown what features would be needed in such a system and whether staff adoption of and engagement with such a system would be successful. The findings presented in this paper provide insights into the architectural design and a detailed list of features for any organization considering implementing a mechanism for handling the critical PROs reported.

In this paper, we present the results of our PRO-based alerting and communication system design, summarize key quantitative results, and reflect on the implications of scaling the adoption of this technology more widely. To our knowledge, this is the first paper to report on the design, implementation, and use of a closed-loop alert management and communication system specifically for managing PRO data in cancer care.

Methods

Ethics Approval

This retrospective cohort study was approved by the Memorial Sloan Kettering (MSK) Cancer Center institutional review board (approval number 19-090) to be conducted between September 2016 and January 2021.

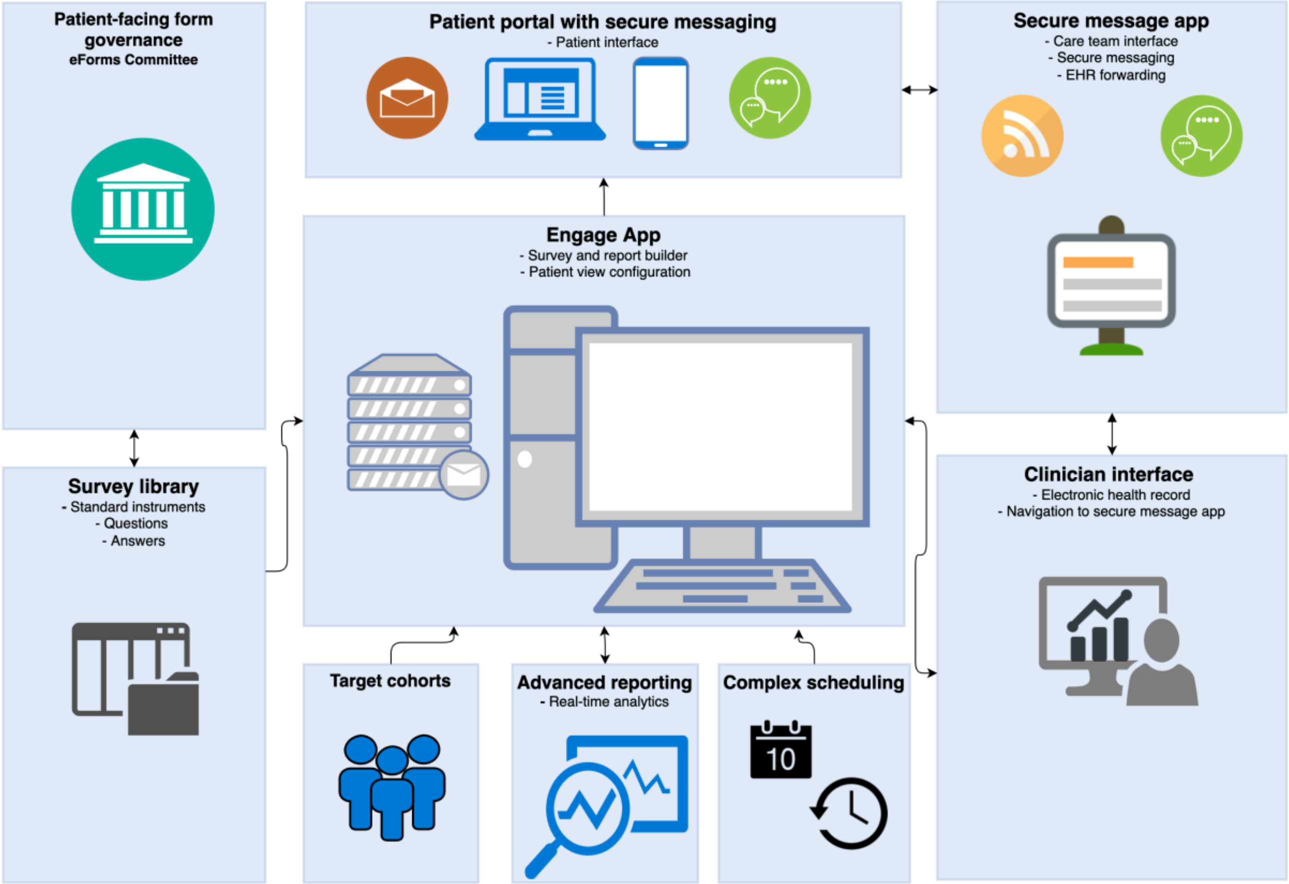
Setting

The study was conducted at a high-volume National Comprehensive Cancer Network in and near the New York City area, across all sites of care, including ambulatory care clinics, inpatient services, ambulatory surgery centers, inpatient surgery, and urgent care. The PRO data collection and alerting system was implemented as a standard of care for multiple cohorts of patients with cancer through individually managed PRO programs consisting of interdisciplinary clinical, administrative, and technical teams. Notably, the novel COVID-19 screening questionnaire and COVID-19 symptom questionnaires were administered to virtually all patients coming to MSK for any appointment. Patients enrolled in these programs would have characteristics similar to those of patients who were more prone to receiving cancer treatments. All numerical results reported were for the entire study period, between September 2016 and January 2021. The median age of the patients was 61 years, and overall instrument compliance was 36.89% (447,562/1,213,271) across all patient cohorts that were part of this study.

Engage System Overview

From September 2016 to August 2017, we launched a pilot where we added alerting features incrementally into Engage, our PRO app, and by rolling out the Recovery Tracker, an electronic postoperative symptom survey based on the PRO–Common Terminology Criteria for Adverse Events [45] that is assigned for 10 days after ambulatory surgery cases. The surveys were completed by patients at home via the patient portal account either through a web browser or our mobile app. Engage, although not the focus of this study, was the foundational backbone for the alert notification and communication system discussed in this study. We launched a total of 86 PRO survey instruments in Engage, including standard of care forms such as intake forms, screening questionnaires, short-term symptom assessments, long-term follow-up questionnaires, and research-based questionnaires. Engage was developed as a stand-alone app; it is tightly integrated within MyMSK (developed internally at MSK patient portal), patient-clinician secure messaging system, and the EHR system [7,46]. Engage is depicted in Figure 1 and consists of 4 key subsystems (survey configuration and deployment, alert notification system, patient and clinician user interfaces, and secure messaging system). Engage is further integrated with upstream databases to support cohort identification and scheduling. It is also integrated with downstream clinical information and documentation systems within the existing clinician and support staff workflows to support clinical charting and escalation workflows. In addition to the technical aspects, the system consists of a governance committee overseeing key design and program decisions called the eForms Committee. The focus of this study is to demonstrate the process behind the design, development, and implementation of the alert notification system and its integrations with upstream (ie, survey library, target cohorts, and complex scheduling) and downstream systems (ie, EHR). The design of Engage (our PRO system) is beyond the scope of this study.

Figure 1. Overview of the homegrown patient-reported outcome (PRO) system called Engage. EHR: electronic health record.



System Governance

To drive decisions and formal governance of the alert notification system, we leveraged our electronic forms committee (eForms Committee), which met monthly. The committee was established at the launch of our PRO initiatives in 2016 to oversee PRO instrument development, evaluate patient burden, discuss impacts on clinician workflow, review regulatory and legal implications, and approve significant changes to the features that were requested by clinical user groups. It is a multidisciplinary committee consisting of health informatics specialists, app development team members, patient engagement specialists, clinicians, researchers, biostatisticians, health information management staff, and hospital administration. In addition to the eForms Committee, PRO work groups were created for each survey instrument, which met more frequently to discuss the management and implementation considerations of the Engage system and provide frequent feedback on system design proposals. PRO work groups also met to decide on setting the initial alerting criteria and adjusting thresholds as needed.

Steps for Alert Notification System Design

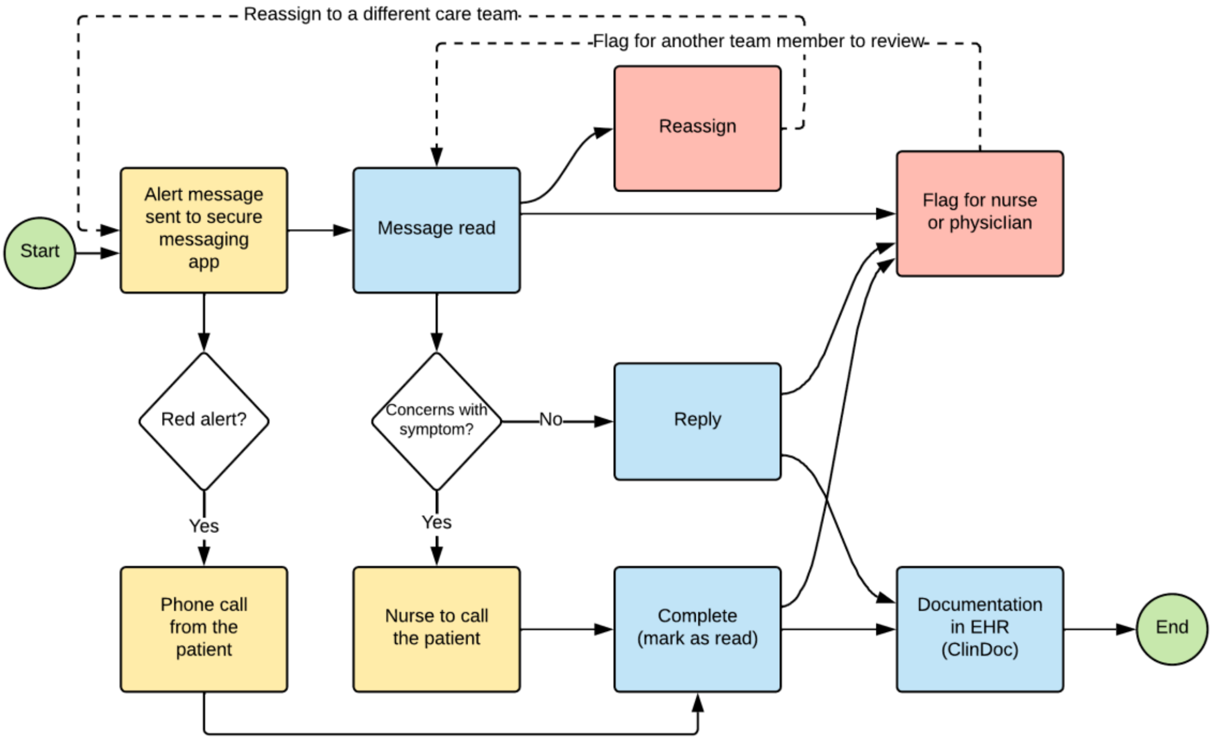
We sought to design and implement a robust and agile PRO-based critical results alert system that notifies the patient and the entire care team of a clinically meaningful patient response as it happens in real time at an oncology care setting. One of the goals for the design was to provide the ability to facilitate secure nonstructured communication between the patient and care team about the abnormal results and to resolve it within the same workflow. A secondary objective of this study was to describe alert volumes, response times, and triaging patterns to understand the implications for scale and feasibility of implementing PROs system wide for all patients with cancer. In addition to the system design and implementation, we created a feedback and governance structure around system enhancements and content revisions based on the Scaled Agile Framework. The feedback was gathered through a series of regular meetings with all project stakeholders including end users, program managers, clinicians, and system developers.

Owing to the lack of an existing methodology for the management of PRO related to critical values (eg, a very severe symptom being reported several days after a surgical event), we referenced models of critical result communication based on abnormal laboratory or radiological findings by reviewing the literature to determine an initial set of desirable system components and features for our PRO-based alert notification system. We conducted a literature review of the existing clinical decision support interventions in PubMed to identify key system components that were necessary to enable clinical alert generation and management. The sample search terms included “critical alert management,” “clinical alert notification,” “critical result notification,” and “abnormal result management.” We identified 5 key capabilities, which were enabled in our alert notification system: alert rule configuration, alert messaging, acknowledgment, triage, and alert export for documentation in the EHR. Once the alerting components were enabled during the pilot, they were adopted by 7 other PRO-based projects, as described in more detail in Multimedia Appendix 1.

Overall, 8 questionnaires were configured with alerting functionalities, targeting more acute symptom assessments, following a recent clinical event that served as a trigger in the target cohorts. In these scenarios, MSK’s best practice expectation was a call back within 2 business days after the clinical event, which would be supplemented with automated symptom assessments. The patients in these cohorts were defined as those who have recently had a surgical event, radiation treatment, chemotherapy treatment, or COVID-19 diagnosis.

Figure 2 shows the alert management workflow followed by the care team members. An alert was defined as a notification that went out to the care team members because of a survey submission by the patient. A patient could report multiple alerting events in surveys that were designed to be recurring for several days (eg, 10 events in a row for a 10-day survey; each survey can result in an alert). In addition, we implemented a patient-facing alert notification, whereby a patient was notified when their responses triggered a concerning symptom via a pop-up on their screen. The notification advised the patient to call their physician’s office if they were concerned about the symptom worsening. At the same time, this triggered an alert message to be sent to their physician’s inbox. Upon reviewing the message, a care team member had the option to call the patient directly to follow-up on any concerning symptoms or reply to the message. Then, they also have the option to flag the message as an escalation indicator for a more senior care team member. Standard nurse phone calls with patients undergoing oncology treatments included questions about any follow-ups after the treatment (eg, symptom assessments and clarifications about PRO responses that may be concerning or need to be elaborated on). Nursing teams also handled triage of any patient concerns as they arose during phone call conversations, including providing patient education materials, facilitating referrals for prescription refills, and referring patients to urgent care facilities or specialty treatment referrals. The decision to call the patient was based on the guidance established by each clinic and the clinical judgment of the care team members. Finally, users have the option to send the message to the EHR to further document it in a clinical note by clicking the “ClinDoc” button.

Figure 2. End-to-end processes for alert message management. EHR: electronic health record.


Reassign to a different care team
Flag for another team member to review
Reassign
Start
Alert message sent to secure messaging app
Message read
Flag for nurse or physician
Red alert?
Concerns with symptom?
No
Reply
Yes
Yes
Phone call from the patient
Nurse to call the patient
Complete (mark as read)
Documentation in EHR (ClinDoc)
End

Applying the Fit Between Individuals, Task, and Technology Framework

During the pilot period between September 2016 and August 2017, we participated in biweekly PRO work group meetings consisting of frontline staff, including physicians, nurses, and advanced practitioners; office staff; and administrative staff to solicit ideas about the system features desired by the care team members reviewing the critical PRO results and communicating with patients. We mapped these features in the Fit between Individuals, Task, and Technology (FITT) framework. We selected the FITT framework because it considers the sociotechnical aspects of a successful system adoption, enabling us to understand the attributes of users, technology, and tasks leading to successful adoption. It also allowed us to consider the interaction of all 3 attribute types to envision a more holistic solution. We continued to use the framework throughout the implementation period to elucidate additional features that were important to consider about the tool, task, and person performing the task for each component of the critical result notification framework identified specifically for PROs [47]. In addition, we discussed workflow aspects of the management of symptom alerts and clinical decision-making processes. Through these meetings, we elucidated the key person and task attributes of designing a PRO-focused alert management system. We presented the model to our biweekly informatics working group consisting of informaticians, system developers, and product managers, where we discussed the task- and tool-related attributes of these capabilities. After each feedback session, we documented features, success criteria, and interventions against the key system capabilities identified in the previous section into a FITT framework, charting them into tool-, task-, and person-related buckets.

Data Collection and Analysis

System use data were collected by querying the underlying reporting database collecting the following variables (defined in Textbox 1) for data between September 2016 and January 2021: patient adoption, patient engagement, alert volume, alert rate, alert types, messaging status, triaging and escalation flags, response times, and clinician involvement. Data were queried using DBeaver software (DBeaver Corporation). Descriptive statistics and data visualizations were developed using Tableau software (Tableau Software, Inc).

Textbox 1. Description of the variables used for analysis.

Variable name and definition

- Patient adoption
  - The total number of patients that completed a patient-reported outcome questionnaire and volume of questionnaires completed
- Patient engagement
  - The percentage of patients who completed a questionnaire out of the total numbers of questionnaires that were assigned per patient
- Alert volume and rate
  - The total number and rate of alerts that were fired per alert-eligible questionnaire
- Alert type
  - The severity level of the alert message (red [severe], red-yellow [severe and moderate], and yellow [moderate]) including unlabeled messages (shown as alert)
- Message status
  - The final status of the alert message in the secure messaging system (read, replied, completed, and documented).
- Message reassignment flag
  - An indicator of whether a message has been reassigned to someone else
- Message escalation flag
  - An indicator of whether a message has been flagged for review by another care team member (either a registered nurse or physician)
- Response time
  - The time between when the alert message was created to when it was last updated by a care team member (in hours)
- Clinician involvement
  - The total number of care teams that were assigned alert messages as measured by unique care team inboxes, including the total number of individual care team members who reviewed, responded to, or handled the alert messages within the care team inboxes

Results

Overview

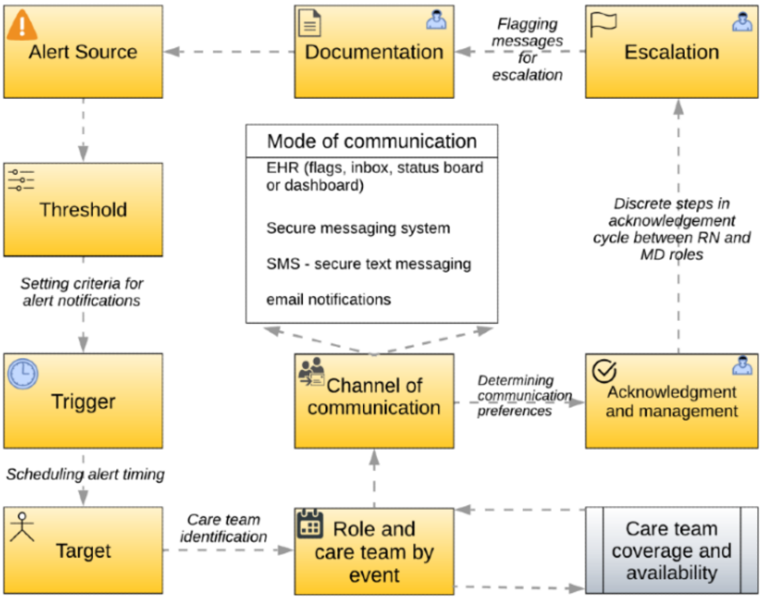
The findings explain how our system was designed, features identified within the FITT framework, how our system creates and schedules alerts, the management and delivery of the alerts, and descriptive statistics of alert management and adoption by the care team members.

Alert Notification System Components and Features

The main system framework components for identifying and communicating critical PRO responses, which we evaluated against the FITT framework and subsequently implemented in our production PRO tool within our patient portal and secure messaging system, are illustrated in Figure 3 and described in Textbox 2.

Figure 3 depicts the overview of the alerting system at a high level. Starting with the source system, Engage, where questionnaires are built, the alerts are configured and patient survey responses are captured and stored. Once a patient submits a questionnaire, the responses are reviewed by a listener to see whether they pass a predefined threshold. Then, when a trigger event specifying the timing of the alert message is detected, the target recipient is identified (this is captured upfront in the patient cohort definition stage based on coverage), and the message is routed to the appropriate communication channel. In our case, this was routed to a secure messaging system, but we configured for an omnichannel communication strategy. The message is sent to a mailbox and reviewed by the care team members, who have the option to escalate it to other care team members or document the conversation in the downstream EHR clinical note.

Figure 3. Patient-reported outcome alert system capabilities and process flow. EHR: electronic health record; MD: Medical Doctor; RN: registered nurse.


Alert Source
Documentation
Flagging messages for escalation
Escalation
Threshold
Mode of communication
EHR (flags, inbox, status board or dashboard)
Secure messaging system
SMS - secure text messaging
email notifications
Discrete steps in acknowledgement cycle between RN and MD roles
Setting criteria for alert notifications
Trigger
Channel of communication
Determining communication preferences
Acknowledgment and management
Scheduling alert timing
Target
Care team identification
Role and care team by event
Care team coverage and availability

Textbox 2. Key components of the alert notification system and their descriptions.

Key component and description

- Alert source and threshold
  - Establish a scoring algorithm based on a single response value or a combination of response values to flag a notification to be sent based on a specific patient submission. One submission can consist of multiple question responses (eg, severe pain and severe fatigue). Responses were presented in sections and color coded to show the most critical alerts first in red, followed by responses that were not critical in yellow so that the care team could triage and prioritize their responses to patients.
- Trigger
  - Define a technical method to schedule the notification to be sent when a patient reports a certain value. Alert messages were stored as JSON objects, and the notification was done in real time for critical alerts.
- Target
  - Define the target system where the message will be visible by the entire patient care team and identify who will receive and manage the message based on their specific clinical role, care relationship with patient, coverage, and availability.
- Communication mode
  - Determine the specific alert communication preferences based on a clinician's role or their tool preference. The communication tools of choice of care team members often varied based on their clinical role. The app accommodated multiple communication modes and the ability to honor the preferred method of communication of each user.
- Acknowledgment and escalation
  - Identify discrete steps in the acknowledgment cycle, including escalation of messages to senior clinical roles for more critical follow-ups. Buttons were created to manage each discrete step in the acknowledgment and escalation processes.
- Documentation
  - Record the most recent status of an alert in the source system, including escalation, or document the follow-up with the patient in the medical record.

FITT Framework Results: Alert Features by Person, Task, and Tool

After we attended biweekly PRO working groups, we charted the desired features and user needs into a modified FITT framework that was stratified by each major component identified in the literature review, as summarized in Textbox 2. This resulted in a comprehensive list of program management processes and system requirements and informed our user acceptance testing scenarios during the development and subsequent rollout of the app. The results are summarized in Table 1.

Table 1. Summarizing capabilities by the tool, task, and person features of our alert system.

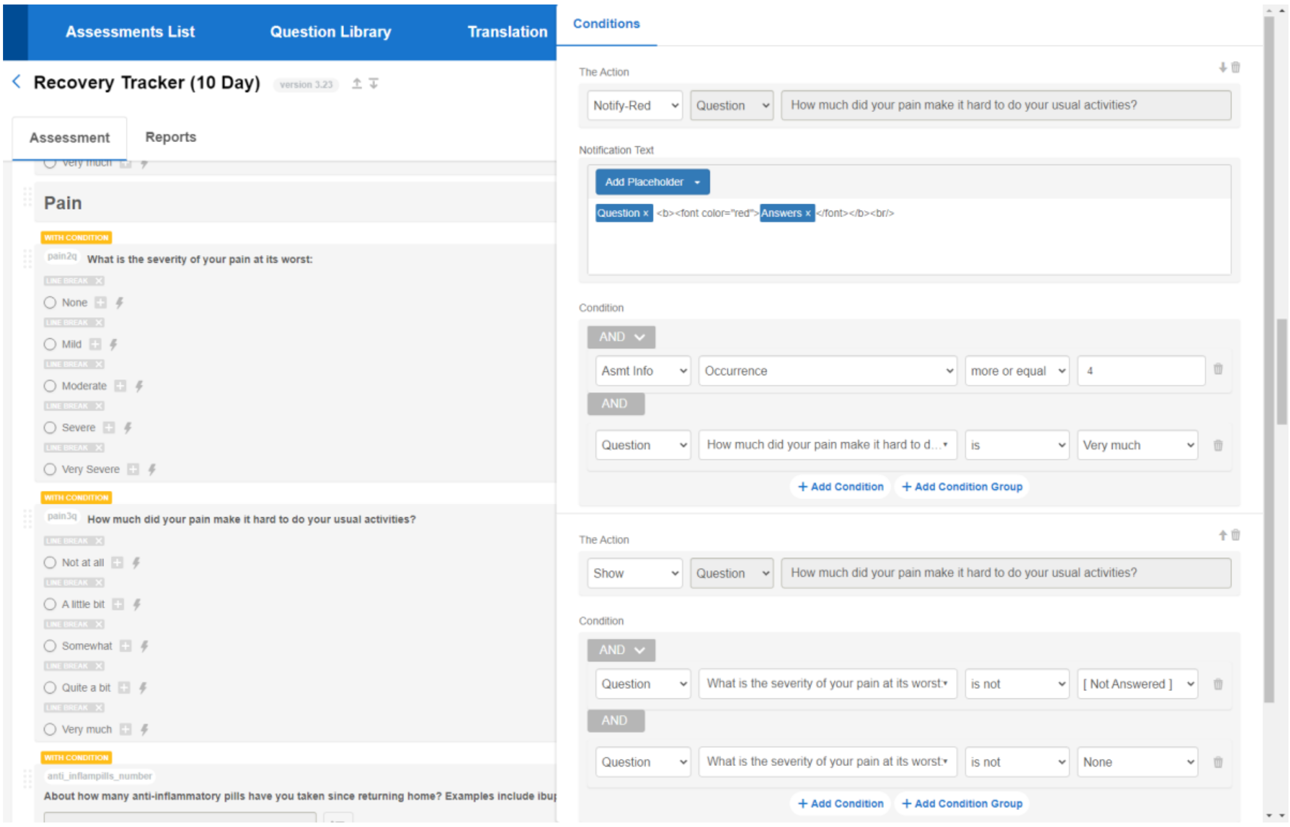
Component	Tool features	Task features	Person features
Alert source authentication	• Care team tracking system in place • Identify systems for generating and receiving notifications • Single sign on to all applications	• Navigation between apps used to respond to an alert • User account verification and management	• Care team identification and authentication to all systems • Patient access to survey submission tool to submit and review patient alerts • Training of users to access and navigate between systems
Alert creation	• Define the customizability points of an alert (frequency, mode, method, and target [person]) • Enable a rules engine to define threshold setting rules and optimization • Bundled alert creation	• Establish trigger points • Identify alert severity levels • Analyze data and tune thresholds for triggers • Define workflows by alert severity (urgent, semiurgent, or nonurgent)	• Governance for creating, reviewing, and tuning alert triggers and thresholds • Patient knowledge of alert creation
Communication	• Interoperable modes of communication established (EHR[a], patient portal, SMS text messaging, apps, email, and telecom [pager, Vocera, telephone, and e-fax]) • Manage preferences for the mode of alert communication	• Redundancy management • Alert bundling and sorting based on similar alerts • Develop definitions of severity language	• Availability and coverage of the care team • Notification preferences established (tool of preference to log into for alerts) by event type or service or patient procedures • Digital communication between the care team and patient
Reminders and escalation	• Method in place to set reminder schedule for critical alerts if they have not been reviewed • Autoescalation of alerts that have not been reviewed	• Rules for reminders • Rules for escalation • Autoescalation • Due date escalation • Missed alert handling	• Department-specific training on acknowledgment management and follow-up actions • Monitoring of escalation patterns
Acknowledgment and management	• Identify systems receiving acknowledgment • Rerouting of messages • Method in place for handling errors in alert creations and communication • Ability to acknowledge a bundle of alerts • Autoacknowledgment	• Define actions that reflect acknowledgment (time, action, and role) • Prioritization based on severity • Voluntary forwarding of alerts • Handling errors in communication	• Training to the care team members on acknowledgment management and follow-up actions • Monitoring of acknowledgment rates
Documentation	• Ability to document alert summary and resolution findings • Ability to copy and paste alert message contents into a clinical note	• Define the levels of documentation to close loop on alert • Feature to import alerts into EHR templates	• Define documentation reviewers • Documentation workflows defined to the care team members, specific to each service and survey

[a] EHR: electronic health record.

Alert Creation and Scheduling

After alert rules had been established by PRO work groups consisting of the most up-to-date clinical standard of practice guidelines adopted by each service, the system administrators were responsible for implementing the criteria. Alert creation was accomplished with a configuration tool in the alert source system, allowing system managers to configure complex rules based on patient responses to individual questions or a combination of questions. The care teams also requested the ability to specify distinct alert rules for specific clinical contexts (such as triggering an alert based on a specific surgical procedure, diagnosis, or treatment regimen) and to vary based on the time span between the clinical event and the time the questionnaire was completed by the patient (such as not firing an alert for pain reported one day after surgery and fire starting after day 3). The alert configuration component, depicted in Figure 4, is where the system administrators configure the subject of the alert message and body of the notification message, including the ability to specify severity levels using visual color indicators and other HTML and cascading style sheets–based text formatting options of the message body. The color label feature was requested by clinicians after spending a few months responding to nonlabeled messages as a mechanism to emphasize severity. Adding color labels that indicate the level of severity to the subject of the message allowed the care team members to triage these notifications appropriately. There was also an ability to integrate the clinical context into the body of the alert notification. Once the alert rules were configured at the questionnaire level, setting a threshold and directionality (greater, equal to, or less than) was the next capability, defined as the level that must be crossed when an alert fires. Next, we developed a triggering mechanism, which is a technical method to configure and synchronize the schedule of sending the notifications based on business rules.

Figure 4. Screenshot of the survey library app, specifically the alert threshold configuration interface.



Recovery Tracker (10 Day)

Pain

pain2q What is the severity of your pain at its worst:

○ None

○ Mild

○ Moderate

○ Severe

○ Very Severe

pain3q How much did your pain make it hard to do your usual activities?

○ Not at all

○ A little bit

○ Somewhat

○ Quite a bit

○ Very much

anti_inflampills_number About how many anti-inflammatory pills have you taken since returning home? Examples include ibu

The Action: Notify-Red | Question | How much did your pain make it hard to do your usual activities?

Notification Text: Question × <b><font color="red"> Answers × </font></b><br/>

Condition: AND
Asmt Info | Occurrence | more or equal | 4
AND
Question | How much did your pain make it hard to d... | is | Very much

The Action: Show | Question | How much did your pain make it hard to do your usual activities?

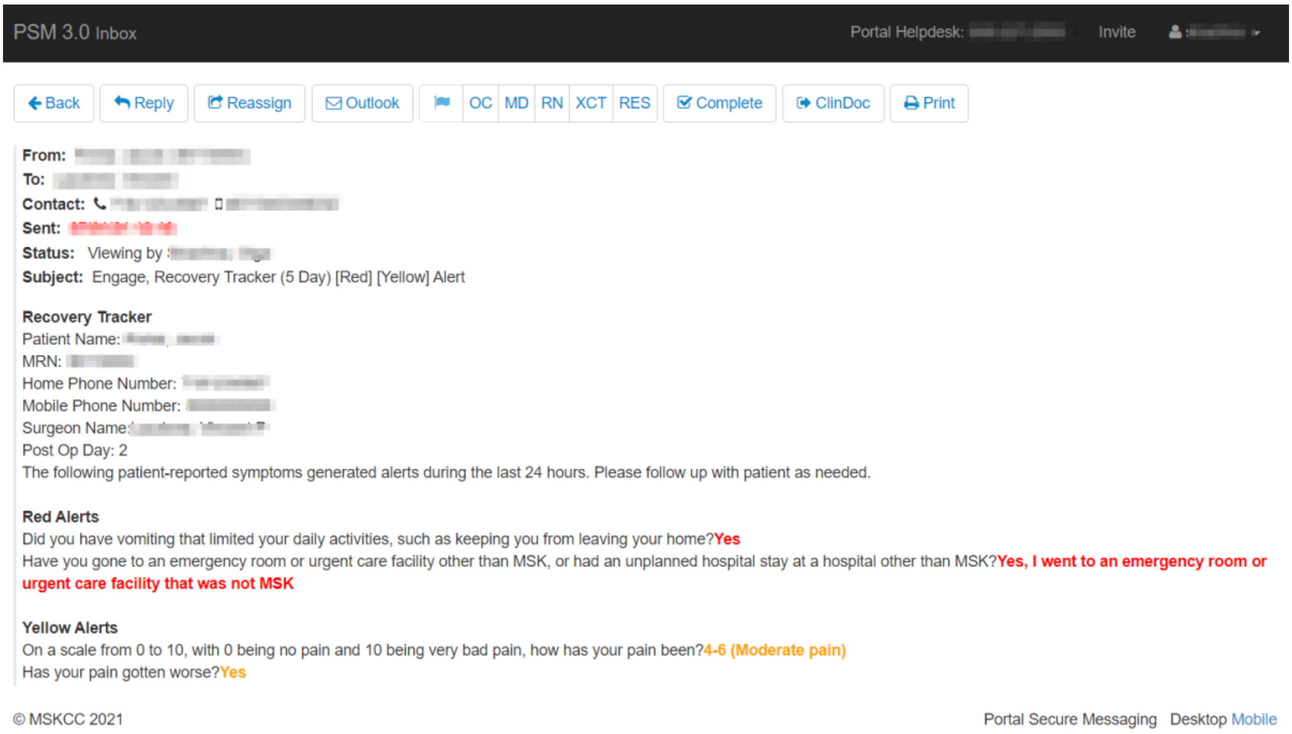
Condition: AND
Question | What is the severity of your pain at its worst: | is not | [Not Answered]
AND
Question | What is the severity of your pain at its worst: | is not | None

Alert Delivery and Management Workflow

Once the alert message is ready to be delivered to a target (defined as a primary clinician taking care of the patient at the time the survey was assigned), there is an ability for the rest of the care team members to subscribe to the primary clinician's secure inbox to review and respond to patient messages on behalf of the entire care team. A screenshot of the alert message is shown in Figure 5. This inbox is built within our patient portal, where the staff can securely communicate with patients bidirectionally. The primary users of the inbox were nursing and administrative office staff supporting the clinic. While in the inbox, users can reassign the message to a different provider if someone else is covering this patient. After opening the message, the staff can acknowledge the message by marking it as complete, reply to the patient directly, or escalate the message to the clinician's office staff, typically a nurse. In addition to the digital workflow, the staff can take the manual route by following up with the patient via a phone call and marking the message as complete. Once an action is taken on the alert message, users have the option to send the message thread to the EHR so that it can be imported into a note. This last import step closes the loop on the alert management life cycle.

Figure 5 demonstrates the output of the alert configuration, which is the message that shows up in the care team's inbox. The message subject indicates the alert severity levels, and the body contains the red or yellow symptom indicators, showing which symptoms triggered the alert. The message body also includes some contextual information about the patient along with their contact information, if available, so that care team members can reach out directly if the message is urgent. The message controls are available on top, supporting the ability to reassign to a different care team member, reply directly to the patient, forward the message to email, flag the message to a different person by role, mark the message as complete, send the message to the EHR (ClinDoc), and finally print the message.

Figure 5. Screenshot of the patient-facing and clinician-facing alert messages. MSK: Memorial Sloan Kettering.



From: [redacted]
To: [redacted]
Contact: ☎ [redacted] [redacted]
Sent: [redacted]
Status: Viewing by [redacted]
Subject: Engage, Recovery Tracker (5 Day) [Red] [Yellow] Alert

Recovery Tracker
Patient Name: [redacted]
MRN: [redacted]
Home Phone Number: [redacted]
Mobile Phone Number: [redacted]
Surgeon Name: [redacted]
Post Op Day: 2
The following patient-reported symptoms generated alerts during the last 24 hours. Please follow up with patient as needed.

Red Alerts
Did you have vomiting that limited your daily activities, such as keeping you from leaving your home?**Yes**
Have you gone to an emergency room or urgent care facility other than MSK, or had an unplanned hospital stay at a hospital other than MSK?**Yes, I went to an emergency room or urgent care facility that was not MSK**

Yellow Alerts
On a scale from 0 to 10, with 0 being no pain and 10 being very bad pain, how has your pain been?**4-6 (Moderate pain)**
Has your pain gotten worse?**Yes**

© MSKCC 2021 Portal Secure Messaging Desktop Mobile

Adoption by Numbers

Through the 8 questionnaires configured with the alerting feature, 7.82% (36,838/470,841) alerts fired per completed symptom assessment, and 13.28% (10,965/82,544) of the patients fired at least one alert out of all patients who received at least one survey. This means that 92.28% (434,003/470,841) of the patient surveys did not trigger an alert; therefore, their responses did not need to be reviewed by the care teams. The alerts were managed by 191 different care teams consisting of 501 staff members. Each care team configured its own mailbox, which was set up at the physician clinic, service, or clinic location level. Staff members subscribe to a care team mailbox and have access to review patient messages, triage them, and respond to the patients. A median of 5 staff members managed each mailbox, with care team volumes ranging from a minimum of 1 to a maximum of 75 staff members. The median number of alerts per care team inbox was 35 (95% CI 0-145).

Alert Management Patterns

Regarding message triage and escalation, 4.4% (1631/36,838) of the alerts were reassigned to other care teams, and 16.7% (6156/36,838) of the alerts were flagged for another care team member to review. The care team members who received the reassigned messages were often located at different campuses, closer to the patients' most recent treatment location rather than the location of the episode that triggered the survey in the first place. In terms of message management, care team members replied to 24.6% (9057/36,838) of the messages and marked 40.9% (15,069/36,838) of the messages as completed (marked as read) without replying, and 34.5% (12,712/36,838) of the messages were further documented in a clinical note within the EHR by following the process shown in Figure 2. Regarding volume, 61.6% (22,692/36,838) of the alerts were yellow, 13.3% (4896/36,838) were red-yellow, 12.4% (4561/36,838) were red, and 12.7% (4689/36,838) were not labeled with a color, which are referred to here as *unlabeled alert*.

When comparing the status of the alert against the alert level, we saw an increase in the care team documentation activity as the alert level increased, which provides a care team–based validation signal of the effectiveness of the alert threshold. The lowest level of alerts being documented were the unlabeled alerts with 27.68% (1298/4689) of all unlabeled alerts documented, whereas the highest level of alerts being documented were the red alerts with 56.83% (2592/4516) of them being documented in the EHR.

We have analyzed the turnaround time to respond to alerts by three different time windows in Table 2: (1) during business hours (8 AM to 6 PM on Monday to Friday), (2) outside business hours (during weekdays between 6 PM and 8 AM the next morning, excluding Sunday to Monday and Friday to Saturday), and (3) over the weekend (after 6 PM on Friday until 8 AM on Monday). The median response time during business hours was 1 hour, with response time varying by alert severity; red alerts had a response time of under an hour, and unlabeled alerts had a median response time of 2 hours. Alerts received outside business hours took longer to review and had a median response time of 6 hours, showing a decrease in response time with increasing severity levels, indicating that care team members used the color label as an effective triaging mechanism, responding to the most critical alerts faster.

Table 2. Summary of alert type by message status and alert arrival window.

	Unlabeled alert (N=4689)	Yellow alert (N=22,692)	Red-yellow alert (N=4896)	Red alert (N=4561)	Total (N=36,838)
Message status, n (%)					
Completed	3026 (64.5)	9718 (42.8)	1630 (33.3)	695 (15.2)	15,069 (40.9)
Replied	365 (7.8)	6040 (26.6)	1378 (28.1)	1274 (27.9)	9057 (24.6)
Documented	1298 (27.7)	6934 (30.6)	1888 (38.6)	2592 (56.8)	12,712 (34.5)
Time window, n (%)					
During business hours	2685 (57.3)	12,159 (53.6)	2597 (53)	2534 (55.6)	19,975 (54.22)
Outside business hours	757 (16.1)	3795 (16.7)	836 (17)	745 (16.3)	6160 (16.72)
Over the weekend	1247 (26.6)	6738 (29.7)	1436 (29.3)	1282 (28.1)	10,703 (29.1)
Time window, median response time in hours					
During business hours	2	1	1	0.5	1
Outside business hours	9	6	7	2	6
Over the weekend	40	23	2	21	22

Discussion

Principal Findings

This study is one of the first studies to report findings on the design, implementation, and operationalization of a PRO critical results management and communication system in cancer care. Although many studies have reported the development of PRO systems, none have focused in depth on the management of results. Our findings suggest that there ought to be a mechanism in place to handle critical patient-reported results in a timely manner so that patients can discuss their symptoms with the care team. To this end, we enhanced an existing secure messaging system to facilitate asynchronous communication between patients and their care teams. Given that the care team can vary in size and composition owing to continually changing shifts, our findings show a median 5-person care team; similar findings have been reported by others [48,49]. It was important to develop a solution where the entire care team had visibility into the prior interactions with the patient to seamlessly pick up the conversation where another care team member left off. The flagging feature was useful for notifying senior team members of a message needing their attention. Although used less frequently, the ability to reassign patients to different care teams allowed for a smooth hand-off between teams.

Setting Clinically Meaningful Alert Thresholds

It was critical to establish clinically meaningful thresholds on a patient cohort, setting a baseline definition of what “normal” symptoms might look like on any given day after a treatment episode. The alert thresholds were a highly debated topic and were revisited many times throughout the post–go-live period of each instrument. The decisions were made within each program work group, where the teams discussed the implications of turning on the alerts and anticipated impacts on workload. The decisions were based on the experience of handling reports of symptoms after treatment episodes targeted for alerting. Clinical care teams consisted of nurses who were well versed in collecting symptom data from patients via phone calls; therefore, they knew which symptoms they would hear on a specific day after an event such as surgery and made their decisions based on clinical judgment. As health care systems learn about patient outcomes over time, it is important to be able to adjust the thresholds. Alert rates were reviewed by staff through summary dashboards, allowing team members to reflect on alert workload burden of staff and determine mitigation strategies. By reviewing the dashboards, management noticed a high alert rate for symptoms such as pain reported the day after a surgical event, which was determined to be a normal clinical event. As a result, alert rules were adjusted to not fire for specific questions within 2 to 3 days after surgery. Setting thresholds such as “red” and “yellow,” which indicate severe and moderate symptoms, respectively, created visual indicators for the care team members within the subject as well as the body of the message. Notably, we saw a substantial difference in response times for the alerts that were marked as “red,” suggesting that the alert color label was effectively used as a triage mechanism. With this approach, by focusing on the alerts that are marked as “red,” we can reduce clinician burnout by minimizing the cognitive load associated with reading patient messages.

Importance of PRO Governance

Similar to other studies on the importance of PRO governance [50], this study also shows that the governance committee (eForms Committee, 35 individuals) was instrumental in the design and implementation of PROs. This committee met monthly to discuss best practices of the overall system design and implications of new feature releases and made decisions establishing clinically meaningful alert thresholds across different patient cohorts. Clinicians, nurses, patient education specialists, and administrative staff served as collaborative thought leaders consistently striving to minimize patient burden and staff alert fatigue through critical assessment of the alert thresholds that were set. The clinically focused PRO work group (40 individuals) met on a more regular biweekly basis to define staff workflows and responsibilities, provide feedback on system design, align alerting with the existing messaging workflows, develop educational material for patients, create training material for staff, and reflect on the summary data presented via real-time dashboards. Policies were established around the roles and responsibilities of the care team members to ensure that messages were responded to in a timely manner. Decidedly, nursing and office staff were instrumental in reviewing responses, triaging, and responding directly to patients, resonating with similar findings of nurse-led patient engagement programs [17,51-54]. With the exception of 1 program, which had dedicated staff reviewing patient messages on weekends, patients were informed that the mailbox was unattended outside of regular business hours and that they were encouraged to call their physician in case of urgent symptoms.

PRO Integration Into Clinical Workflows

Integrating PROs into clinical workflows has always been a challenge, and having clinicians review and act on the data is yet another challenge. Establishing an alerting system, notifying clinicians of only clinically meaningful patient responses via alerts is a step toward a better direction, where clinicians only need to review alerts if they are deemed clinically significant by clinical expert consensus. Given the emerging problem of clinician burnout, partially caused by information overload, by using this alerting system, the care teams reviewed 7.82% (36,838/470,841) of the patient responses and eliminated the need to review over 92.28% (434,003/470,841) of the responses, while maintaining the collection of valuable PRO data to study long-term patient outcomes in response to treatments. Having a seamless mechanism in place to communicate with patients within the same workflow is yet another step in the right direction and is aligned with findings from the literature [27]. Not only does it signal to the patient that they are being constantly cared for, encouraging them to keep completing their assessments, but it can also be used to address the symptoms early, preventing any unnecessary emergency room visits.

Future Research Opportunities

Although PRO data can be a valuable tool for shared decision-making and bridging the care gap for in-between visits, the data are only available if patients complete their assessments. The adoption of remote monitoring programs during the COVID-19 pandemic [55-57] exposed the digital divide created by programs solely relying on digital interventions. Patients who are not as comfortable with technology or those whose primary language is different from the language of the survey instrument may be less likely to complete their symptom assessments. As we scale PRO-based remote symptom monitoring programs, we must consider studying the sociotechnical aspects of a wholly digital intervention. There are several implications for future research opportunities with respect to setting meaningful alert thresholds appropriately. At MSK, the deliberations around setting clinically meaningful thresholds evolved over time and, ultimately, were decided through agreement between nursing staff accountable for responding to alerts and physicians responsible for patient outcomes. In addition, it would be interesting to analyze the impact of patient characteristics, such as demographics, disease stage, or disease type, on alert response patterns and communication with care team members. As we accumulate more robust PRO data sets and monitor clinician triage, there is an opportunity to build machine learning models to predict when patients will need interventions based on their responses to specific PROs, response patterns, and clinical context such as disease stage and progression. In addition, automated artificial intelligence–based chatbots can be developed to facilitate conversations with patients, reducing the burden on nursing staff. Further studying follow-up activities of nurses in EHRs such as referrals, medication orders, or communication with other care team members can inform the refinement in the chatbot responses to patients.

Limitations

This study has some limitations. First, we report findings from one institution, which may not be generalizable across all settings. In addition, because we have an in-house–developed patient engagement system, we had the flexibility to design and implement an alert management system that was best suited to the care team workflows, which may not be a flexibility affordable to other health care institutions. Moreover, the care teams that opted to implement alerting features for their PRO programs were highly engaged in the system development life cycle and provided ample feedback throughout the process, which may not apply to institutions with limited resources. System design features for most of the functionalities were informed through discussions with a pilot work group and were qualitative in nature, and we did not perform a formal quantitative assessment.

Conclusions

By developing a critical symptom alerting and communication system, we designed a system supporting the real-time delivery of critical results based on PRO data to appropriate care team members, including the ability for a patient and clinical staff to communicate in a nonstructured, text-based, secure communication format about the alert. We were able to standardize the processing of patient-generated alert messages, enabling the presentation of clinically meaningful PRO data within clinical workflows in a standard format, and monitor response times by clinical staff. This allowed us to set an appropriate patient expectation for a response time frame by their care team members or provided alternate communication guidance specific to each patient and the surgical procedure they underwent.

Acknowledgments

The authors would like to thank the members of the eForms Committee and Josie Robertson Surgery Center patient-reported outcome work group for their constant and consistent feedback throughout the product development life cycle, from the inception phase until the present day. The authors would also like to acknowledge the efforts of the nursing, informatics, and technical teams at Memorial Sloan Kettering who contributed to the development, implementation, and support of the Engage platform and the alerting notification system.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of questionnaires using the alerting features implemented in this study.
[DOCX File , 14 KB - medinform_v10i11e38483_app1.docx]

References

1. Velikova G, Booth L, Smith AB, Brown PM, Lynch P, Brown JM, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. J Clin Oncol 2004 Mar 15;22(4):714-724. [doi: 10.1200/JCO.2004.06.078] [Medline: 14966096]
2. Basch E, Mody GN, Dueck AC. Electronic patient-reported outcomes as digital therapeutics to improve cancer outcomes. JCO Oncol Pract 2020 Sep;16(9):541-542. [doi: 10.1200/OP.20.00264] [Medline: 32484724]
3. Chung AE, Basch EM. Incorporating the patient's voice into electronic health records through patient-reported outcomes as the "review of systems". J Am Med Inform Assoc 2015 Jul;22(4):914-916 [FREE Full text] [doi: 10.1093/jamia/ocu007] [Medline: 25614143]
4. Howell D, Li M, Sutradhar R, Gu S, Iqbal J, O'Brien MA, et al. Integration of patient-reported outcomes (PROs) for personalized symptom management in "real-world" oncology practices: a population-based cohort comparison study of impact on healthcare utilization. Support Care Cancer 2020 Oct;28(10):4933-4942. [doi: 10.1007/s00520-020-05313-3] [Medline: 32020357]
5. Cowan RA, Suidan RS, Andikyan V, Rezk YA, Einstein MH, Chang K, et al. Electronic patient-reported outcomes from home in patients recovering from major gynecologic cancer surgery: a prospective study measuring symptoms and health-related quality of life. Gynecol Oncol 2016 Nov;143(2):362-366 [FREE Full text] [doi: 10.1016/j.ygyno.2016.08.335] [Medline: 27637366]
6. Yang LY, Manhas DS, Howard AF, Olson RA. Patient-reported outcome use in oncology: a systematic review of the impact on patient-clinician communication. Support Care Cancer 2018 Jan;26(1):41-60. [doi: 10.1007/s00520-017-3865-7] [Medline: 28849277]
7. Strachna O, Cohen MA, Allison MM, Pfister DG, Lee NY, Wong RJ, et al. Case study of the integration of electronic patient-reported outcomes as standard of care in a head and neck oncology practice: obstacles and opportunities. Cancer 2021 Feb 01;127(3):359-371 [FREE Full text] [doi: 10.1002/cncr.33272] [Medline: 33107986]
8. Richards HS, Blazeby JM, Portal A, Harding R, Reed T, Lander T, et al. A real-time electronic symptom monitoring system for patients after discharge following surgery: a pilot study in cancer-related surgery. BMC Cancer 2020 Jun 10;20(1):543 [FREE Full text] [doi: 10.1186/s12885-020-07027-5] [Medline: 32522163]
9. Warrington L, Absolom K, Conner M, Kellar I, Clayton B, Ayres M, et al. Electronic systems for patients to report and manage side effects of cancer treatment: systematic review. J Med Internet Res 2019 Jan 24;21(1):e10875 [FREE Full text] [doi: 10.2196/10875] [Medline: 30679145]
10. Bennett AV, Jensen RE, Basch E. Electronic patient-reported outcome systems in oncology clinical practice. CA Cancer J Clin 2012;62(5):337-347 [FREE Full text] [doi: 10.3322/caac.21150] [Medline: 22811342]
11. Basch E, Stover AM, Schrag D, Chung A, Jansen J, Henson S, et al. Clinical utility and user perceptions of a digital system for electronic patient-reported symptom monitoring during routine cancer care: findings from the PRO-TECT trial. JCO Clin Cancer Inform 2020 Oct;4:947-957 [FREE Full text] [doi: 10.1200/CCI.20.00081] [Medline: 33112661]
12. Basch E, Artz D, Iasonos A, Speakman J, Shannon K, Lin K, et al. Evaluation of an online platform for cancer patient self-reporting of chemotherapy toxicities. J Am Med Inform Assoc 2007;14(3):264-268 [FREE Full text] [doi: 10.1197/jamia.M2177] [Medline: 17329732]
13. Basch E, Abernethy AP. Supporting clinical practice decisions with real-time patient-reported outcomes. J Clin Oncol 2011 Mar 10;29(8):954-956. [doi: 10.1200/JCO.2010.33.2668] [Medline: 21282536]
14. Anatchkova M, Donelson SM, Skalicky AM, McHorney CA, Jagun D, Whiteley J. Exploring the implementation of patient-reported outcome measures in cancer care: need for more real-world evidence results in the peer reviewed literature. J Patient Rep Outcomes 2018 Dec 27;2(1):64 [FREE Full text] [doi: 10.1186/s41687-018-0091-0] [Medline: 30588562]
15. Pugh SL, Rodgers JP, Moughan J, Bonanni R, Boparai J, Chen RC, et al. Do reminder emails and past due notifications improve patient completion and institutional data submission for patient-reported outcome measures? Qual Life Res 2021 Jan;30(1):81-89 [FREE Full text] [doi: 10.1007/s11136-020-02613-3] [Medline: 32894431]
16. Stover AM, Basch EM. Using patient-reported outcome measures as quality indicators in routine cancer care. Cancer 2016 Mar 01;122(3):355-357 [FREE Full text] [doi: 10.1002/cncr.29768] [Medline: 26619153]

17. Stover AM, Tompkins Stricker C, Hammelef K, Henson S, Carr P, Jansen J, et al. Using stakeholder engagement to overcome barriers to implementing patient-reported outcomes (PROs) in cancer care delivery: approaches from 3 prospective studies. Med Care 2019 May;57 Suppl 5 Suppl 1:S92-S99. [doi: 10.1097/MLR.0000000000001103] [Medline: 30985602]
18. Holt JM, Cusatis R, Winn A, Asan O, Spanbauer C, Williams JS, et al. Impact of pre-visit contextual data collection on patient-physician communication and patient activation: a randomized trial. J Gen Intern Med 2021 Nov;36(11):3321-3329. [doi: 10.1007/s11606-020-06583-7] [Medline: 33559067]
19. Hsiao CJ, Dymek C, Kim B, Russell B. Advancing the use of patient-reported outcomes in practice: understanding challenges, opportunities, and the potential of health information technology. Qual Life Res 2019 Jun;28(6):1575-1583. [doi: 10.1007/s11136-019-02112-0] [Medline: 30684149]
20. Holch P, Warrington L, Bamforth LC, Keding A, Ziegler LE, Absolom K, et al. Development of an integrated electronic platform for patient self-report and management of adverse events during cancer treatment. Ann Oncol 2017 Sep 01;28(9):2305-2311 [FREE Full text] [doi: 10.1093/annonc/mdx317] [Medline: 28911065]
21. Schwartzberg L. Electronic patient-reported outcomes: the time is ripe for integration into patient care and clinical research. Am Soc Clin Oncol Educ Book 2016 May 19(36):e89-e96 [FREE Full text] [doi: 10.1200/edbk_158749]
22. Cusatis R, Holt JM, Williams J, Nukuna S, Asan O, Flynn KE, et al. The impact of patient-generated contextual data on communication in clinical practice: a qualitative assessment of patient and clinician perspectives. Patient Educ Couns 2020 Apr;103(4):734-740. [doi: 10.1016/j.pec.2019.10.020] [Medline: 31744702]
23. Austin E, LeRouge C, Hartzler AL, Chung AE, Segal C, Lavallee DC. Opportunities and challenges to advance the use of electronic patient-reported outcomes in clinical care: a report from AMIA workshop proceedings. JAMIA Open 2019 Dec;2(4):407-410 [FREE Full text] [doi: 10.1093/jamiaopen/ooz042] [Medline: 32025635]
24. Pitzen C, Larson J. Patient-reported outcome measures and integration into electronic health records. J Oncol Pract 2016 Oct;12(10):867-872. [doi: 10.1200/JOP.2016.014118] [Medline: 27460494]
25. Dalal AK, Pesterev BM, Eibensteiner K, Newmark LP, Samal L, Rothschild JM. Linking acknowledgement to action: closing the loop on non-urgent, clinically significant test results in the electronic health record. J Am Med Inform Assoc 2015 Jul;22(4):905-908 [FREE Full text] [doi: 10.1093/jamia/ocv007] [Medline: 25796594]
26. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, with the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Apr 10;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
27. Ye J. The impact of electronic health record-integrated patient-generated health data on clinician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):1051-1056 [FREE Full text] [doi: 10.1093/jamia/ocab017] [Medline: 33822095]
28. Edrees H, Amato MG, Wong A, Seger DL, Bates DW. High-priority drug-drug interaction clinical decision support overrides in a newly implemented commercial computerized provider order-entry system: override appropriateness and adverse drug events. J Am Med Inform Assoc 2020 Jun 01;27(6):893-900 [FREE Full text] [doi: 10.1093/jamia/ocaa034] [Medline: 32337561]
29. Wright A, McEvoy DS, Aaron S, McCoy AB, Amato MG, Kim H, et al. Structured override reasons for drug-drug interaction alerts in electronic health records. J Am Med Inform Assoc 2019 Oct 01;26(10):934-942 [FREE Full text] [doi: 10.1093/jamia/ocz033] [Medline: 31329891]
30. Absolom K, Holch P, Warrington L, Samy F, Hulme C, Hewison J, eRAPID systemic treatment work group. Electronic patient self-Reporting of Adverse-events: Patient Information and aDvice (eRAPID): a randomised controlled trial in systemic cancer treatment. BMC Cancer 2017 May 08;17(1):318 [FREE Full text] [doi: 10.1186/s12885-017-3303-8] [Medline: 28482877]
31. Bates DW, Teich JM, Lee J, Seger D, Kuperman GJ, Ma'Luf N, et al. The impact of computerized physician order entry on medication error prevention. J Am Med Inform Assoc 1999;6(4):313-321 [FREE Full text] [doi: 10.1136/jamia.1999.00660313] [Medline: 10428004]
32. Al-Mutairi A, Meyer AN, Chang P, Singh H. Lack of timely follow-up of abnormal imaging results and radiologists' recommendations. J Am Coll Radiol 2015 Apr;12(4):385-389. [doi: 10.1016/j.jacr.2014.09.031] [Medline: 25582812]
33. Singh H, Thomas EJ, Sittig DF, Wilson L, Espadas D, Khan MM, et al. Notification of abnormal lab test results in an electronic medical record: do any safety concerns remain? Am J Med 2010 Mar;123(3):238-244 [FREE Full text] [doi: 10.1016/j.amjmed.2009.07.027] [Medline: 20193832]
34. Singh H, Wilson L, Reis B, Sawhney MK, Espadas D, Sittig DF. Ten strategies to improve management of abnormal test result alerts in the electronic health record. J Patient Saf 2010 Jun;6(2):121-123 [FREE Full text] [doi: 10.1097/PTS.0b013e3181ddf652] [Medline: 20563228]
35. Singh H, Thomas EJ, Mani S, Sittig D, Arora H, Espadas D, et al. Timely follow-up of abnormal diagnostic imaging test results in an outpatient setting: are electronic medical records achieving their potential? Arch Intern Med 2009 Sep 28;169(17):1578-1586 [FREE Full text] [doi: 10.1001/archinternmed.2009.263] [Medline: 19786677]
36. Lacson R, O'Connor SD, Andriole KP, Prevedello LM, Khorasani R. Automated critical test result notification system: architecture, design, and assessment of provider satisfaction. AJR Am J Roentgenol 2014 Nov;203(5):W491-W496 [FREE Full text] [doi: 10.2214/AJR.14.13063] [Medline: 25341163]

37. Middleton B, Sittig DF, Wright A. Clinical decision support: a 25 year retrospective and a 25 year vision. Yearb Med Inform 2016 Aug 02;Suppl 1:S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]
38. Hassett MJ, Hazard H, Osarogiagbon RU, Wong SL, Bian JJ, Dizon DS, eSyM Project Managers. Design of eSyM: an ePRO-based symptom management tool fully integrated in the electronic health record (Epic) to foster patient/clinician engagement, sustainability, and clinical impact. J Clin Oncol 2020 May 25;38(15_suppl):e14120. [doi: 10.1200/jco.2020.38.15_suppl.e14120]
39. Basch E, Deal AM, Dueck AC, Scher HI, Kris MG, Hudis C, et al. Overall survival results of a trial assessing patient-reported outcomes for symptom monitoring during routine cancer treatment. JAMA 2017 Jul 11;318(2):197-198 [FREE Full text] [doi: 10.1001/jama.2017.7156] [Medline: 28586821]
40. Avery KN, Richards HS, Portal A, Reed T, Harding R, Carter R, et al. Developing a real-time electronic symptom monitoring system for patients after discharge following cancer-related surgery. BMC Cancer 2019 May 17;19(1):463 [FREE Full text] [doi: 10.1186/s12885-019-5657-6] [Medline: 31101017]
41. Annis T, Pleasants S, Hultman G, Lindemann E, Thompson JA, Billecke S, et al. Rapid implementation of a COVID-19 remote patient monitoring program. J Am Med Inform Assoc 2020 Aug 01;27(8):1326-1330 [FREE Full text] [doi: 10.1093/jamia/ocaa097] [Medline: 32392280]
42. Taylor PC. Adopting PROs in virtual and outpatient management of RA. Nat Rev Rheumatol 2020 Sep;16(9):477-478 [FREE Full text] [doi: 10.1038/s41584-020-0449-6] [Medline: 32504074]
43. Aiyegbusi OL, Calvert MJ. Patient-reported outcomes: central to the management of COVID-19. Lancet 2020 Aug 22;396(10250):531 [FREE Full text] [doi: 10.1016/S0140-6736(20)31724-4] [Medline: 32791038]
44. Marandino L, Necchi A, Aglietta M, Di Maio M. COVID-19 emergency and the need to speed up the adoption of electronic patient-reported outcomes in cancer clinical practice. JCO Oncol Pract 2020 Jun;16(6):295-298 [FREE Full text] [doi: 10.1200/OP.20.00237] [Medline: 32364846]
45. Basch E, Reeve BB, Mitchell SA, Clauser SB, Minasian LM, Dueck AC, et al. Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). J Natl Cancer Inst 2014 Sep;106(9):dju244 [FREE Full text] [doi: 10.1093/jnci/dju244] [Medline: 25265940]
46. Polubriaginof FC, Parekh PK, Akella NR, Stetson PD. Adoption patterns of an electronic patient-reported outcomes tool in oncology. J Clin Oncol 2020 May 25;38(15_suppl):e19127. [doi: 10.1200/jco.2020.38.15_suppl.e19127]
47. Ammenwerth E, Iller C, Mahler C. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. BMC Med Inform Decis Mak 2006 Jan 09;6:3 [FREE Full text] [doi: 10.1186/1472-6947-6-3] [Medline: 16401336]
48. Steitz BD, Unertl KM, Levy MA. Characterizing communication patterns among members of the clinical care team to deliver breast cancer treatment. J Am Med Inform Assoc 2020 Feb 01;27(2):236-243 [FREE Full text] [doi: 10.1093/jamia/ocz151] [Medline: 31682267]
49. Steitz BD, Unertl KM, Levy MA. An analysis of electronic health record work to manage asynchronous clinical messages among breast cancer care teams. Appl Clin Inform 2021 Aug;12(4):877-887 [FREE Full text] [doi: 10.1055/s-0041-1735257] [Medline: 34528233]
50. Stetson PD, McCleary NJ, Osterman T, Ramchandran K, Tevaarwerk A, Wong T, et al. Adoption of patient-generated health data in oncology: a report from the NCCN EHR oncology advisory group. J Natl Compr Canc Netw (forthcoming) 2022 Jan 18:1-6. [doi: 10.6004/jnccn.2021.7088] [Medline: 35042190]
51. Simon BA, Assel MJ, Tin AL, Desai P, Stabile C, Baron RH, et al. Association between electronic patient symptom reporting with alerts and potentially avoidable urgent care visits after ambulatory cancer surgery. JAMA Surg 2021 Aug 01;156(8):740-746 [FREE Full text] [doi: 10.1001/jamasurg.2021.1798] [Medline: 34076691]
52. Nembhard IM, Buta E, Lee YS, Anderson D, Zlateva I, Cleary PD. A quasi-experiment assessing the six-months effects of a nurse care coordination program on patient care experiences and clinician teamwork in community health centers. BMC Health Serv Res 2020 Mar 24;20(1):137 [FREE Full text] [doi: 10.1186/s12913-020-4986-0] [Medline: 32093664]
53. Epstein AS, Desai AV, Bernal C, Romano D, Wan PJ, Okpako M, et al. Giving voice to patient values throughout cancer: a novel nurse-led intervention. J Pain Symptom Manage 2019 Jul;58(1):72-9.e2 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.04.028] [Medline: 31034869]
54. Desai AV, Klimek VM, Chow K, Epstein AS, Bernal C, Anderson K, et al. 1-2-3 project: a quality improvement initiative to normalize and systematize palliative care for all patients with cancer in the outpatient clinic setting. J Oncol Pract 2018 Dec;14(12):e775-e785 [FREE Full text] [doi: 10.1200/JOP.18.00346] [Medline: 30537456]
55. Sirintrapun SJ, Lopez AM. Telemedicine in cancer care. Am Soc Clin Oncol Educ Book 2018 May 23;38:540-545 [FREE Full text] [doi: 10.1200/EDBK_200141] [Medline: 30231354]
56. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. J Am Med Inform Assoc 2020 Jul 01;27(7):1147-1148 [FREE Full text] [doi: 10.1093/jamia/ocaa078] [Medline: 32343813]
57. Bakhtiar M, Elbuluk N, Lipoff JB. The digital divide: how COVID-19's telemedicine expansion could exacerbate disparities. J Am Acad Dermatol 2020 Nov;83(5):e345-e346 [FREE Full text] [doi: 10.1016/j.jaad.2020.07.043] [Medline: 32682890]

Abbreviations

EHR: electronic health record
FITT: Fit between Individuals, Task, and Technology
MSK: Memorial Sloan Kettering
PRO: patient-reported outcome

Edited by C Lovis; submitted 04.04.22; peer-reviewed by M Zaidi, F Sparano, FM Calisto, M Marie; comments to author 18.05.22; revised version received 27.06.22; accepted 25.09.22; published 03.11.22.

Please cite as:
Strachna O, Asan O, Stetson PD
Managing Critical Patient-Reported Outcome Measures in Oncology Settings: System Development and Retrospective Study
JMIR Med Inform 2022;10(11):e38483
URL: https://medinform.jmir.org/2022/11/e38483
doi:10.2196/38483
PMID:36326801

Original Paper

Training a Deep Contextualized Language Model for International Classification of Diseases, 10th Revision Classification via Federated Learning: Model Development and Validation Study

Pei-Fu Chen[1,2*], MD; Tai-Liang He[3*], MSc; Sheng-Che Lin[3], MSc; Yuan-Chia Chu[4,5,6], PhD; Chen-Tsung Kuo[4,5,6], PhD; Feipei Lai[1,3,7], PhD; Ssu-Ming Wang[1], MSc; Wan-Xuan Zhu[8], BSc; Kuan-Chih Chen[1,9], MSc, MD; Lu-Cheng Kuo[10], MSc, MD; Fang-Ming Hung[11,12], MD; Yu-Cheng Lin[13,14], MD, PhD; I-Chang Tsai[15], PhD; Chi-Hao Chiu[16], MS; Shu-Chih Chang[17], MA; Chi-Yu Yang[18,19], MD

[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]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[4]Department of Information Management, Taipei Veterans General Hospital, Taipei City, Taiwan
[5]Medical Artificial Intelligence Development Center, Taipei Veterans General Hospital, Taipei City, Taiwan
[6]Department of Information Management, National Taipei University of Nursing and Health Sciences, Taipei City, Taiwan
[7]Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[8]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan
[9]Department of Internal Medicine, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[10]Department of Internal Medicine, National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan
[11]Department of Medical Affairs, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[12]Department of Surgical Intensive Care Unit, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[13]Department of Pediatrics, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[14]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan
[15]Artificial Intelligence Center, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[16]Section of Health Insurance, Department of Medical Affairs, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[17]Medical Records Department, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[18]Department of Information Technology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
[19]Section of Cardiovascular Medicine, Cardiovascular Center, Far Eastern Memorial Hospital, New Taipei City, Taiwan
*these authors contributed equally

Corresponding Author:
Chi-Yu Yang, MD
Department of Information Technology
Far Eastern Memorial Hospital
No 21, Section 2, Nanya S Rd, Banciao District
New Taipei City, 220216
Taiwan
Phone: 886 2 8966 7000
Email: chiyuyang1959@gmail.com

Abstract

Background: The automatic coding of clinical text documents by using the *International Classification of Diseases, 10th Revision* (ICD-10) can be performed for statistical analyses and reimbursements. With the development of natural language processing models, new transformer architectures with attention mechanisms have outperformed previous models. Although multicenter training may increase a model's performance and external validity, the privacy of clinical documents should be protected. We used federated learning to train a model with multicenter data, without sharing data per se.

Objective: This study aims to train a classification model via federated learning for ICD-10 multilabel classification.

Methods: Text data from discharge notes in electronic medical records were collected from the following three medical centers: Far Eastern Memorial Hospital, National Taiwan University Hospital, and Taipei Veterans General Hospital. After comparing the performance of different variants of bidirectional encoder representations from transformers (BERT), PubMedBERT was chosen for the word embeddings. With regard to preprocessing, the nonalphanumeric characters were retained because the model's performance decreased after the removal of these characters. To explain the outputs of our model, we added a label attention mechanism to the model architecture. The model was trained with data from each of the three hospitals separately and via federated learning. The models trained via federated learning and the models trained with local data were compared on a testing set that was composed of data from the three hospitals. The micro F_1 score was used to evaluate model performance across all 3 centers.

Results: The F_1 scores of PubMedBERT, RoBERTa (Robustly Optimized BERT Pretraining Approach), ClinicalBERT, and BioBERT (BERT for Biomedical Text Mining) were 0.735, 0.692, 0.711, and 0.721, respectively. The F_1 score of the model that retained nonalphanumeric characters was 0.8120, whereas the F_1 score after removing these characters was 0.7875—a decrease of 0.0245 (3.11%). The F_1 scores on the testing set were 0.6142, 0.4472, 0.5353, and 0.2522 for the federated learning, Far Eastern Memorial Hospital, National Taiwan University Hospital, and Taipei Veterans General Hospital models, respectively. The explainable predictions were displayed with highlighted input words via the label attention architecture.

Conclusions: Federated learning was used to train the ICD-10 classification model on multicenter clinical text while protecting data privacy. The model's performance was better than that of models that were trained locally.

(JMIR Med Inform 2022;10(11):e41342) doi:10.2196/41342

KEYWORDS

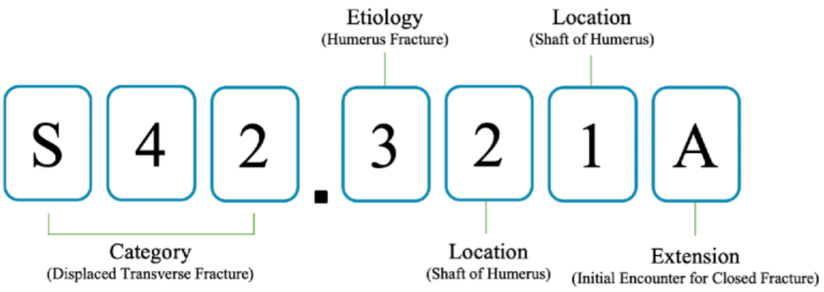
federated learning; International Classification of Diseases; machine learning; natural language processing; multilabel text classification

Introduction

Background

The World Health Organization published a unified classification system for diagnoses of diseases called the *International Classification of Diseases* (ICD), and the ICD 10th Revision (ICD-10) is widely used [1]. Coders classify diseases according to the rules of the ICD, and the resulting ICD codes are used for surveys, statistics, and reimbursements. The ICD-10 Clinical Modification (ICD-10-CM) is used for coding medical diagnoses and includes approximately 69,000 codes [2,3]. ICD-10-CM codes contain 7 digits; the structure is shown in Figure 1.

Figure 1. Structure of an *International Classification of Diseases, 10th Revision, Clinical Modification* code.



In hospitals, diagnoses for each patient are first written as text descriptions in the electronic health record. A coder then reads these records to classify diagnoses into ICD codes. Because diagnoses are initially written as free text, the text's ambiguity makes diagnoses difficult to code. Classifying each diagnosis is very time-consuming. A discharge record may contain 1 to 20 codes. Per the estimation of a trial, coders spent 20 minutes assigning codes to each patient on average [4]. An automatic tool can be used to increase the efficiency of and reduce the labor for ICD classification.

Related Work

Recently, deep learning and natural language processing (NLP) models have been developed to turn plain text into vectors, making it possible to automatically classify them. Shi et al [5] proposed a hierarchical deep learning model with an attention mechanism. Sammani et al [6] introduced a bidirectional gated recurrent unit model to predict the first 3 or 4 digits of ICD codes based on discharge letters. Wang et al [7] proposed a convolutional neural network model with an attention mechanism and gated residual network to classify Chinese records into ICD codes. Makohon et al [8] showed that deep learning with an attention mechanism effectively enhances ICD-10 predictions. Previous studies also mentioned the necessity of enormous data sets and how privacy-sensitive clinical data limited the development of models for automatic ICD-10 classification [6].

Federated learning has achieved impressive results in the medical field, being used to train models on multicenter data while keeping them private. Federated learning is widely used in medical image and signal analyses, such as brain imaging analysis [9] and the classification of electroencephalography signals [10]. In the clinical NLP field, Liu et al [11] proposed a 2-stage federated method that involved using clinical notes from different hospitals to extract phenotypes for medical tasks.

Previously, we applied a Word2Vec model with a bidirectional gated recurrent unit to classify ICD-10-CM codes from electronic medical records [12]. We analyzed the distribution of ICD-10-CM codes and extracted features from discharge notes. The model had an F_1 score of 0.625 for ICD-10-CM code classification. To improve the model's performance, we implemented bidirectional encoder representations from transformers (BERT) and found an improved F_1 score of 0.715 for ICD-10-CM code classification [4]. We also found that the coding time decreased when coders used classification model aids; the median F_1 score significantly improved from 0.832 to 0.922 (P<.05) in a trial [4]. Furthermore, we constructed a system to improve ease of use, comprising data processing, feature extraction, model construction, model training, and a web service interface [4]. Lastly, we included a rule-based algorithm in the preprocessing process and improved the F_1 score to 0.853 for ICD-10-CM classification [13].

Objective

This study aims to further improve the performance of the ICD-10 classification model and enable the model's use across hospitals. In this study, we investigated the effect of federated learning on the performance of a model that was trained on medical text requiring ICD-10 classification.

Methods

Ethics Approval

The study protocol was approved by the institutional review boards of Far Eastern Memorial Hospital (FEMH; approval number: 109086-F), National Taiwan University Hospital (NTUH; approval number: 201709015RINC), and Taipei Veterans General Hospital (VGHTPE; approval number: 2022-11-005AC), and the study adhered to the tenets of the Declaration of Helsinki. Informed consent was not applicable due to the use of deidentified data.

Data Collection

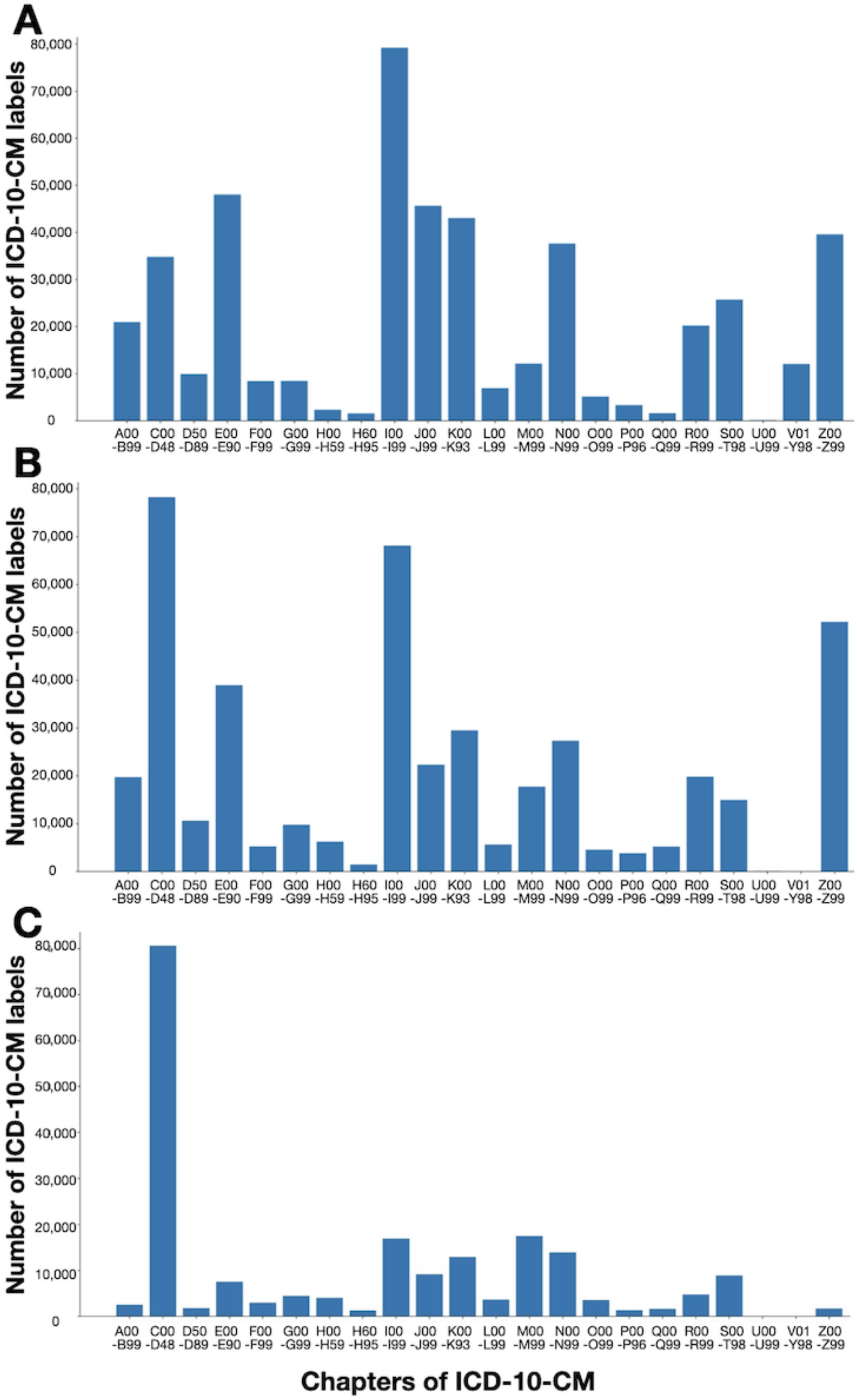
Our data were acquired from electronic health records at FEMH (data recorded between January 2018 and December 2020), NTUH (data recorded between January 2016 and July 2018), and VGHTPE (data recorded between January 2018 and December 2020). The data contained the text of discharge notes and ICD-10-CM codes. Coders in each hospital annotated the ground truth ICD-10 codes.

Data Description

After duplicate records were removed, our data set contained 100,334, 239,592, and 283,535 discharge notes from FEMH, NTUH, and VGHTPE, respectively. Each record contained between 1 and 20 ICD-10-CM labels. The distribution of labels for each chapter is shown in Figure 2. These chapters are classified by the first three digits. Codes for chapters V01 to Y98 are not used for insurance reimbursement; hence, they were excluded from our data set. The minimum number of ICD-10-CM labels was found for chapters U00 to U99, and the maximum number was found for chapters J00 to J99. Counts of ICD-10-CM labels from the three hospitals are shown in Multimedia Appendix 1.

The text in the data set contained alphabetic characters, punctuation, and a few Chinese characters. The punctuation count and the top 10 Chinese characters are shown in Multimedia Appendix 2. The most common punctuation mark was the period ("."), and the least common was the closing brace ("}").

Figure 2. Counts of ICD-10-CM labels for 22 chapters from (A) Far Eastern Memorial Hospital, (B) National Taiwan University Hospital, and (C) Taipei Veterans General Hospital. ICD-10-CM: *International Classification of Diseases, 10th Revision, Clinical Modification*.



Preprocessing

We first removed duplicate medical records from the data set. We then transformed all full-width characters into half-width characters and all alphabetic characters into lowercase letters. Records shorter than 5 characters were removed, as these were usually meaningless words, such as “nil” and “none.” We also removed meaningless characters, such as newlines, carriage returns, horizontal tabs, and formed characters (“\n,” “\r,” “\t,” and “\f,” respectively). Finally, all text fields were concatenated.

To choose a better method for managing punctuation and Chinese characters during the preprocessing stage, we determined model performance by using FEMH data, given the inclusion of these characters in the data. Each experiment used 2 versions of the data. In the first version, we retained these specific characters, and in the second, we removed them.

Experiment P investigated the effect of punctuation, experiment C investigated the effect of Chinese characters, and experiment PC investigated the effects of both punctuation and Chinese characters. Another method of retaining Chinese character information is using English translations of Chinese characters. Therefore, we also compared the model's performance when Chinese characters were retained to its performance when Google Translate was used to obtain English translations.

One-hot encoding was used for the labels. Of the 69,823 available ICD-10-CM codes, 17,745 appeared in our combined data set, resulting in a one-hot encoding vector length of 17,745. The final cohort comprised 100,334, 239,592, and 283,535 records from FEMH, NTUH, and VGHTPE, respectively; 20% (FEMH: 20,067/100,334; NTUH: 47,918/239,592; VGHTPE: 56,707/283,535) of the records were randomly selected for the testing set, and the remaining records were used as the training set.

Classification Model

We compared the performance of different variants of BERT, including PubMedBERT [14], RoBERTa (Robustly Optimized BERT Pretraining Approach) [15], ClinicalBERT [16], and BioBERT (BERT for Biomedical Text Mining) [17]. BioBERT was pretrained with text from PubMed—the most popular bibliographic database in the health and medical science fields. ClinicalBERT was pretrained with the MIMIC-III (Medical Information Mart for Intensive Care III) data set, and its vocabulary was from English Wikipedia and the BookCorpus data set. PubMedBERT is another variant of BERT that uses training data from PubMed. The main difference between PubMedBERT and BioBERT is their vocabularies. The vocabulary of BioBERT was from English Wikipedia and the BookCorpus data set—as was the vocabulary of BERT—whereas that of PubMedBERT was from PubMed. This difference in vocabularies affects the ability to recognize words in clinical text. RoBERTa used the original BERT model, but it also used a longer training time, a larger batch size, and more training data. The training data were from the BookCorpus, CC-News (CommonCrawl News), and OpenWebText data sets. RoBERTa also applied dynamic masking, which meant that the masked tokens would be changed multiple times instead of being fixed in the original BERT. The vocabularies and corpora of these BERT variants are summarized in Table 1.

For our comparison, the text was first fed into the BERT tokenizer, which transformed strings into tokens. The number of tokens was then truncated to 512 for every text datum that met the input length limit of 512. A linear layer connected the word embeddings produced from the models to the output layers of the one-hot–encoded multilabels. The output size of the linear layer was 17,745, which matched the one-hot encoding vector size of the labels. Binary cross-entropy was used to calculate the model loss. We trained our model for 100 epochs, with a learning rate of 0.00005. These models were fine-tuned for our ICD-10-CM multilabel classification task to compare their performance. Figure 3 summarizes the model architecture and preprocessing flowchart. The best-performing model and preprocessing method were chosen for subsequent federated learning.

Table 1. Summary of the vocabulary and corpus sources for the various bidirectional encoder representations from transformers (BERT) models.

Models	Vocabulary sources	Corpus sources (training data)
PubMedBERT	PubMed	PubMed
RoBERTa[a]	The BookCorpus, CC-News[b], and OpenWebText data sets	The BookCorpus, CC-News, and OpenWebText data sets
ClinicalBERT	English Wikipedia and the BookCorpus data set	The MIMIC-III[c] data set
BioBERT[d]	English Wikipedia and the BookCorpus data set	PubMed

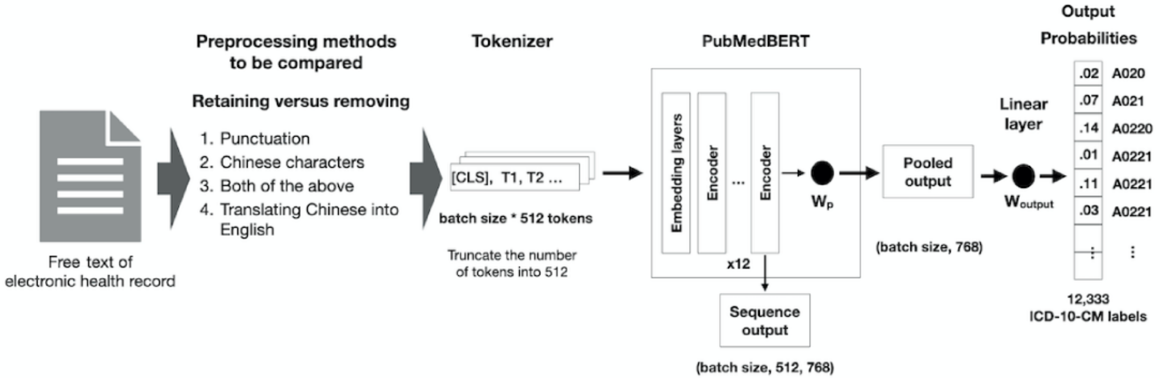
[a] RoBERTa: Robustly Optimized BERT Pretraining Approach.

[b] CC-News: CommonCrawl News.

[c] MIMIC-III: Medical Information Mart for Intensive Care III.

[d] BioBERT: BERT for Biomedical Text Mining.

Figure 3. Model architecture and processing flowchart. CLS: class token; ICD-10-CM: *International Classification of Diseases, 10th Revision, Clinical Modification*.



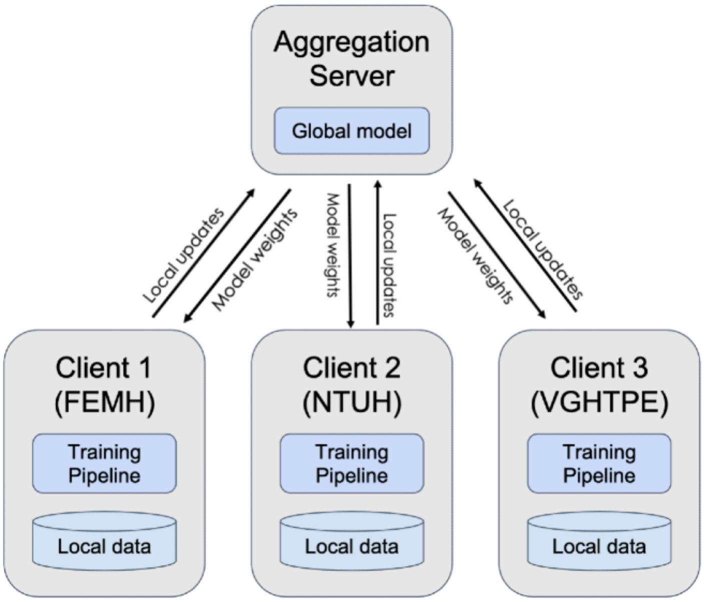
Federated Learning

With federated learning, a model can be trained without sharing data [18]. Clients (ie, local machines) keep their training data on the same model architecture while exchanging the weights of model parameters. A server receives the weights from each client and averages their weights. After updating the model, the server sends new weights back to the clients. The clients can then start a new training round. We updated the weights of our model parameters with the *FederatedAveraging* algorithm [18] and used Flower for federated learning [19].

Flower is an open-source federated learning framework for researchers [19]. Flower has a server-client structure. The server and clients need to be started individually, and a server needs to be assigned to each client. They communicate via the open-source Google Remote Procedure Call (gRPC; Google LLC) [20]. With the gRPC, a client application can directly call a method on a server application, and this can be done on different machines. There is a registration center on the server for managing communication with all clients. There are 3 main modules in the server. The first—a connection management module—maintains all current gRPC connections. On the server, each gRPC corresponds to each client. When a gRPC is established, the register function is triggered to store the clients' information in an array. If a client initiates a disconnection or the connection times out, the register function will be called to clear the client. The second module—a bridge module—caches the information, regardless of whether the gRPC information from the clients or the server will be stored in the module. However, since the buffer is shared in both directions, it is necessary to use the state transition method to ensure that all of the information in the buffer is the same. There are five states—the *close*, *waiting for client write*, *waiting for client read*, *waiting for server write*, and *waiting for server read* states. The third module—a server handler—manages the traffic between the server and the clients.

Clients were set in the three hospitals, where the model was trained on local data. The weights from each client were transferred to the server, where the weights were averaged, and global models were made (Figure 4). We set 5 epochs for each training round on clients and 20 rounds for the server aggregation. Our study was conducted on 2 nodes. Each node had a NVIDIA RTX 2080 Ti graphics processing unit (NVIDIA Corporation) with 64 GB of RAM, and one node had 2 NVIDIA TITAN RTX graphics processing units with 64 GB of RAM (NVIDIA Corporation).

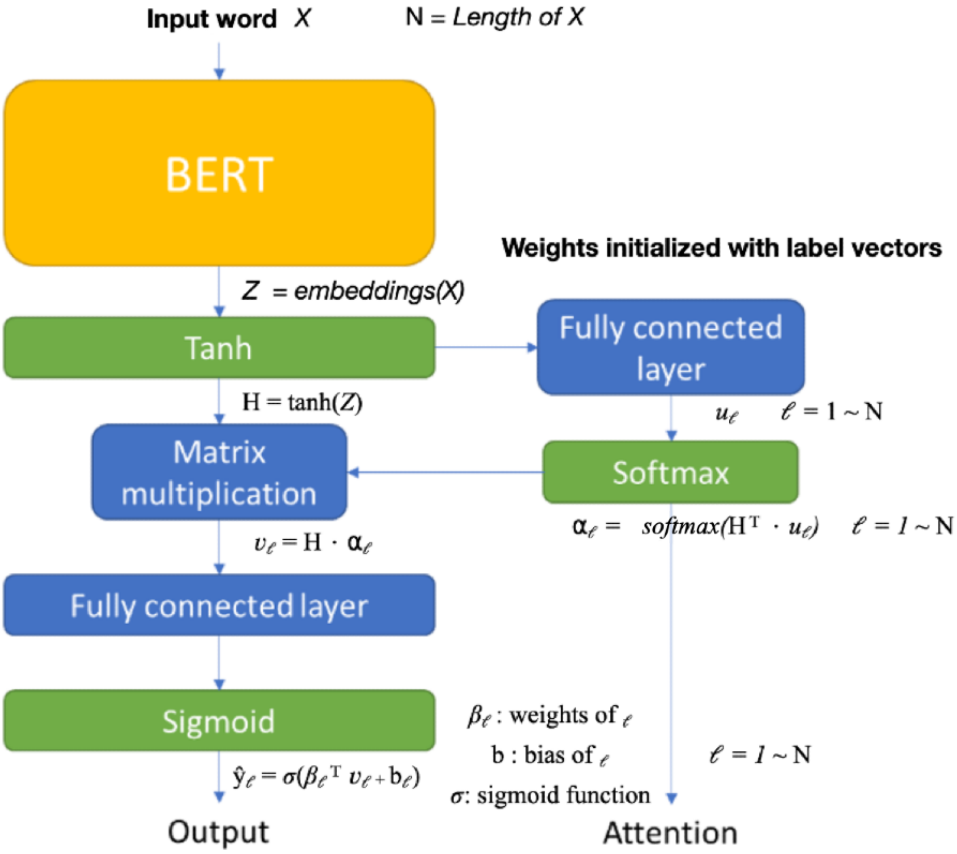
Figure 4. Federated learning architecture. FEMH: Far Eastern Memorial Hospital; NTUH: National Taiwan University Hospital; VGHTPE: Taipei Veterans General Hospital.



Label Attention

To explain the outputs of our model, we added a label attention architecture [21]. It calculated the attention based on the inner products of word vectors and each label vector separately. Figure 5 shows how we added the label attention architecture to our model. First, we fine-tuned the BERT model by using the definitions of ICD-10-CM codes to generate the label vectors. Second, we constructed a fully connected layer, of which the weights were initialized with the label vectors. Third, the output produced by BERT was passed through the hyperbolic tangent function, thereby producing word vectors. We inputted the word vectors (Z) into the fully connected layer and softmax layer. The output ( ) of the softmax layer was the attention. Fourth, we inputted the hyperbolic tangent function of word vectors (H), which were multiplied by attention ( ), into another fully connected layer and sigmoid layer. This was similar to our original architecture. The output (y) could be subtracted from the one-hot–encoded labels for the loss calculation. Finally, attention was used to explain how the model predicted the labels. Attention was given to the input text for corresponding ICD-10-CM codes. The performance of the model after adding the label attention architecture was compared to its performance without this architecture.

Figure 5. Our model architecture with label attention. BERT: bidirectional encoder representations from transformers.


Input word X
N = *Length of X*
BERT
Weights initialized with label vectors
Z = *embeddings*(X)
Tanh
Fully connected layer
$\mathrm{H} = \tanh(Z)$
u_ℓ $\ell = 1 \sim \mathrm{N}$
Matrix multiplication
Softmax
$\alpha_\ell = softmax(\mathrm{H}^\mathrm{T} \cdot u_\ell)$ $\ell = 1 \sim \mathrm{N}$
$v_\ell = \mathrm{H} \cdot \alpha_\ell$
Fully connected layer
Sigmoid
β_ℓ : weights of ℓ
b : bias of ℓ
$\ell = 1 \sim \mathrm{N}$
$\hat{y}_\ell = \sigma(\beta_\ell{}^\mathrm{T}\, v_\ell + \mathrm{b}_\ell)$
σ: sigmoid function
Output
Attention

Metrics

We used the micro F_1 score to evaluate performance because it is the harmonic mean of precision and recall and therefore yields more balanced results than those yielded when using precision or recall only. The micro F_1 score was calculated as follows:

where

and

TP_{sum} indicates the sum of true positives, FP_{sum} indicates the sum of false positives, and FN_{sum} indicates the sum of false negatives.

Results

Comparing the Performance of Different BERT Models

The F_1 scores of PubMedBERT, RoBERTa, ClinicalBERT, and BioBERT were 0.735, 0.692, 0.711, and 0.721, respectively. The F_1 score of PubMedBERT was the highest, and that of RoBERTa was the lowest among all models (Table 2). Due to these results, we used PubMedBERT in the subsequent experiments.

Table 2. Performance of different bidirectional encoder representations from transformers (BERT) models.

Models	F_1 score	Precision	Recall
PubMedBERT	0.735	0.756	0.715
RoBERTa[a]	0.692	0.719	0.666
ClinicalBERT	0.711	0.735	0.689
BioBERT[b]	0.721	0.754	0.691

[a] RoBERTa: Robustly Optimized BERT Pretraining Approach.

[b] BioBERT: BERT for Biomedical Text Mining.

The Model's Performance When Retaining or Removing Punctuation or Chinese Characters

Table 3 shows the mean number of tokens for each data set preprocessing case. The mean number of tokens when removing punctuation and Chinese characters was 52.9. The mean number of tokens when the characters were retained in experiment P (punctuation), experiment C (Chinese characters), and experiment PC (punctuation and Chinese characters) was 65.0, 53.1, and 65.1, respectively. Punctuation and Chinese characters comprised 18.3% (1,301,988/7,096,460) and 0.1% (7948/7,096,460) of the tokens in our data, respectively.

Table 3. Mean number of data tokens for retaining or removing punctuation or Chinese characters.

Experiment	Mean number of tokens
Removed punctuation and Chinese characters (baseline)	52.9
Retained punctuation	65.0
Retained Chinese characters	53.1
Retained punctuation and Chinese characters	65.1

Table 4 shows the F_1 scores for each data set preprocessing case. The baseline performance of the model after removing punctuation and Chinese characters was 0.7875. In experiment P, the F_1 score for retaining punctuation was 0.8049—an increase of 0.0174 (2.21%). In experiment C, the F_1 score for retaining Chinese characters was 0.7984—an increase of 0.0109 (1.38%). In experiment PC, the F_1 score for retaining punctuation and Chinese characters was 0.8120—an increase of 0.0245 (3.11%). In all experiments, retaining these characters was better than removing them, with experiment PC showing the largest improvement in performance.

Table 4. F_1 scores for retaining or removing punctuation or Chinese characters.

Experiment	F_1 score	Absolute increases (percentage)
Removed punctuation and Chinese characters (baseline)	0.7875	N/A[a]
Retained punctuation	0.8049	0.0174 (2.21%)
Retained Chinese characters	0.7984	0.0109 (1.38%)
Retained punctuation and Chinese characters	0.8120	0.0245 (3.11%)

[a] N/A: not applicable.

The Model's Performance Before and After Translation

In the experiment where we translated Chinese into English, the F_1 score for retaining the Chinese characters was 0.7984, and that for translating them into English was 0.7983.

Federated Learning

Table 5 shows the performance of the models that were trained in the three hospitals. The models trained in FEMH, NTUH, and VGHTPE had validation F_1 scores of 0.7802, 0.7718, and 0.6151, respectively. The FEMH model had testing F_1 scores of 0.7412, 0.5116, and 0.1596 on the FEMH, NTUH, and VGHTPE data sets, respectively. The NTUH model had testing F_1 scores of 0.5583, 0.7710, and 0.1592 on the FEMH, NTUH, and VGHTPE data sets, respectively. The VGHTPE model had testing F_1 scores of 0.1081, 0.1058, and 0.5692 on the FEMH, NTUH, and VGHTPE data sets, respectively. The weighted average testing F_1 scores were 0.4472, 0.5353, and 0.2522 for the FEMH, NTUH, and VGHTPE models, respectively.

Table 6 shows the federated learning model's performance in the three hospitals. The federated learning model had validation F_1 scores of 0.7464, 0.6511, and 0.5979 on the FEMH, NTUH, and VGHTPE data sets, respectively. The federated learning model had testing F_1 scores of 0.7103, 0.6135, and 0.5536 on the FEMH, NTUH, and VGHTPE data sets, respectively. The weighted average testing F_1 score was 0.6142 for the federated learning model.

Table 5. Models that were trained in the three hospitals for *International Classification of Diseases, 10th Revision* classification.

Hospitals	Validation F_1 score	Testing F_1 scores	Weighted average testing F_1 scores
FEMH[a]	0.7802	• 0.7412 (FEMH) • 0.5116 (NTUH[b]) • 0.1596 (VGHTPE[c])	0.4472
NTUH	0.7718	• 0.5583 (FEMH) • 0.7710 (NTUH) • 0.1592 (VGHTPE)	0.5353
VGHTPE	0.6151	• 0.1081 (FEMH) • 0.1058 (NTUH) • 0.5692 (VGHTPE)	0.2522

[a] FEMH: Far Eastern Memorial Hospital.

[b] NTUH: National Taiwan University Hospital.

[c] VGHTPE: Taipei Veterans General Hospital.

Table 6. The federated learning model's performance in the three hospitals.

Data	Validation F_1 score	Testing F_1 score[a]
FEMH[b] data	0.7464	0.7103
NTUH[c] data	0.6511	0.6135
VGHTPE[d] data	0.5979	0.5536

[a] The weighted average testing F_1 score was 0.6142.

[b] FEMH: Far Eastern Memorial Hospital.

[c] NTUH: National Taiwan University Hospital.

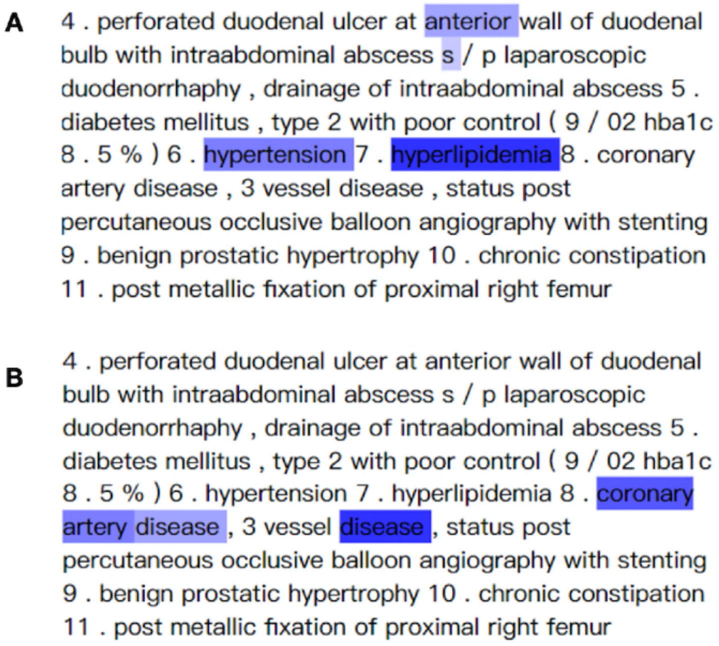
[d] VGHTPE: Taipei Veterans General Hospital.

Label Attention

The F_1 scores of the model with and without the label attention mechanism were 0.804 (precision=0.849; recall=0.763) and 0.813 (precision=0.852; recall=0.777), respectively.

Figure 6 shows a visualization of the attention for ICD-10-CM codes and their related input text. The words were colored blue based on the attention scores for different labels. The intensity of the blue color represented the magnitude of the attention score. We used ICD-10-CM codes E78.5 (“Hyperlipidemia, unspecified”) and I25.10 (“Atherosclerotic heart disease of native coronary artery without angina pectoris”) as examples.

Figure 6. Attention for *International Classification of Diseases, 10th Revision, Clinical Modification* codes (A) E78.5 ("Hyperlipidemia, unspecified") and (B) I25.10 ("Atherosclerotic heart disease of native coronary artery without angina pectoris"). The intensity of the blue color represents the magnitude of the attention score.



A

4 . perforated duodenal ulcer at anterior wall of duodenal bulb with intraabdominal abscess s / p laparoscopic duodenorrhaphy , drainage of intraabdominal abscess 5 . diabetes mellitus , type 2 with poor control (9 / 02 hba1c 8 . 5 %) 6 . hypertension 7 . hyperlipidemia 8 . coronary artery disease , 3 vessel disease , status post percutaneous occlusive balloon angiography with stenting 9 . benign prostatic hypertrophy 10 . chronic constipation 11 . post metallic fixation of proximal right femur

B

4 . perforated duodenal ulcer at anterior wall of duodenal bulb with intraabdominal abscess s / p laparoscopic duodenorrhaphy , drainage of intraabdominal abscess 5 . diabetes mellitus , type 2 with poor control (9 / 02 hba1c 8 . 5 %) 6 . hypertension 7 . hyperlipidemia 8 . coronary artery disease , 3 vessel disease , status post percutaneous occlusive balloon angiography with stenting 9 . benign prostatic hypertrophy 10 . chronic constipation 11 . post metallic fixation of proximal right femur

Discussion

Principal Findings

The federated learning model outperformed each local model when tested on external data. The weighted average F_1 scores on the testing set were 0.6142, 0.4472, 0.5353, and 0.2522 for the federated learning, FEMH, NTUH, and VGHTPE models, respectively (Table 5 and Table 6). The model's performance decreased when tested on external data. Because different doctors, coders, and diseases are found in different hospitals, the style of clinical notes may be distinct across hospitals. Overcoming such gaps among hospitals is challenging. Although the performance of the federated learning model was inferior to that of the models trained on local data when tested on local data, its performance was higher than that of the models trained on local data when tested on external data. Moreover, in the VGHTPE data set, the label distribution was very different from the label distributions in the other two hospitals' data sets (Figure 2). Therefore, the VGHTPE model only achieved F_1 scores of 0.1058 and 0.1081 on the NTUH and FEMH testing sets, respectively. The FEMH and NTUH models had F_1 scores of 0.1596 and 0.1592, respectively, on the VGHTPE testing set (Table 5).

Federated learning improves model performance on external data. Federated learning can be used to build an ICD coding system for use across hospitals. However, the training time required for federated learning is longer than the training time required for local deep learning. Federated learning takes approximately 1 week, and local training takes approximately 2 days. There are 2 reasons for this. First, the communication between the server and the clients takes longer if the model is large. The size of our model is approximately 859 MB. Second, different clients may have different computing powers, and the slower client becomes a bottleneck [22,23]. Other clients may wait for the slower client until it completes its work.

The performance of PubMedBERT was better than that of BioBERT, ClinicalBERT, and RoBERTa. Table 2 shows that the vocabulary of BERT models is an important factor of model performance. The vocabulary of PubMedBERT contains predominantly medical terms, whereas the vocabularies of the other three models contain common words. This difference affects the ability to recognize words in clinical text. Most published BERT models use a vocabulary of 30,522 WordPieces [24]. However, these vocabulary data do not contain some words from special fields. For example, the medical term "lymphoma" is in the vocabulary of PubMedBERT but not in the vocabularies of BioBERT, ClinicalBERT, and RoBERTa. The term "lymphoma" can be transformed into the token "lymphoma" by the PubMedBERT tokenizer, but the term would be split into 3 tokens—"l", "##ymph", and "##oma"—by BioBERT, ClinicalBERT, and RoBERTa.

In most scenarios, nonalphanumeric characters are removed because they are considered useless to the models [25]. In contrast to models with attention mechanisms, early NLP models could not pay attention to punctuation. Additional characters would make the models unable to focus well on keywords. The removal of punctuation in English text and text in other languages, such as Arabic, has been performed for NLP [26]. Ek et al [27] compared 2 data sets of daily conversation text—one retained punctuation, and the other did not. Their results showed better performance for the data set that retained punctuation.

For experiments P, C, and PC, all models performed better when additional characters were retained (Table 4). Experiment P demonstrated that PubMedBERT could use embedded punctuation. As punctuation marks are used to separate different sentences, removing them connects all sentences and thus makes it harder for a model to understand the text content. The improvement in our F_1 score for retaining punctuation is similar to the results of previous work by Ek et al [27]. Our results demonstrate that retaining punctuation can improve the performance of text classification models for text from the clinical field. Experiment C demonstrated that PubMedBERT could use embedded Chinese characters. Although PubMedBERT was pretrained with mostly English text, its vocabulary contains many Chinese characters. The tokens from Chinese characters may contribute to the ICD-10 classification task for clinical text because they provide information such as place names, trauma mechanisms, and local customs. The results of experiment PC indicate that the benefits of retaining punctuation and retaining Chinese characters are additive. In the translation experiment, the F_1 scores did not considerably differ. This result indicates that the model can extract information from clinical text in either English or Chinese. The use of the attention mechanisms of BERT increased our model's ability to pay attention to keywords. Punctuation and Chinese characters contribute helpful information to these models. Therefore, this preprocessing strategy—retaining more meaningful tokens—provides more information for ICD-10 classification task models.

In our previous study, we introduced an attention mechanism to visualize the attention given to the input text for ICD-10 definitions [4]. Through this approach, we trained a model to predict ICD-10 codes and trained another model to extract attention data. This approach might result in inconsistencies between the predictions and attention. In this study, we introduced the label attention architecture to visualize the attention given to the input text for ICD-10 codes [21]. This method better illustrated the attention given to the input words that were used to predict ICD codes, as it is consistent with the methods used by prediction models.

The F_1 score of the model, after the label attention mechanism was added, decreased by 0.009. Although the F_1 score decreased, we obtained explainable predictions. For ICD-10-CM codes E78.5 ("Hyperlipidemia, unspecified") and I25.10 ("Atherosclerotic heart disease of native coronary artery without angina pectoris"), our model successfully paid great attention to the related words "hyperlipidemia" and "coronary artery" (Figure 6). Our visualization method (ie, highlighting input words) allows users to understand how our model identified ICD-10-CM codes from text.

Limitations

Our study has several limitations. First, our data were acquired from 3 tertiary hospitals in Taiwan. The extrapolation of our results to hospitals in other areas should be studied in the future. Second, although our results suggest that model performance is better when punctuation and Chinese characters are retained, this effect may be restricted to specific note types. This finding should be further examined in the context of classifying other types of clinical text. Third, the translated text in our last experiment may not be as accurate as translations by a native speaker. However, it is difficult to manually translate large amounts of data. As such, we could only automatically translate the text by using Google Translate.

It should be noted that there is a primary and secondary diagnosis code for each discharge note. Although choosing the primary code makes reimbursements different, the model proposed in this study did not identify primary codes. To make our model capable of identifying a primary code, we proposed a sequence-to-sequence model in our previous work [4]. It transforms the original predicted labels that were concatenated alphabetically, so that they are ordered by diagnosis code. This structure can be added to the model proposed in this study. Predictions based on primary and secondary diagnosis codes can further improve the usability of this system.

Conclusions

Federated learning was used to train the ICD-10 classification model on multicenter clinical text while protecting data privacy. The model's performance was better than that of models that were trained locally. We showed the explainable predictions by highlighting input words via a label attention architecture. We also found that the PubMedBERT model can use the meanings of punctuation and non-English characters. This finding demonstrates that changing the preprocessing method for ICD-10 multilabel classification can improve model performance.

Acknowledgments

This study was supported by grants from the Ministry of Science and Technology, Taiwan (grants MOST 110-2320-B-075-004-MY and MOST 110-2634-F-002-032-); Far Eastern Memorial Hospital, Taiwan (grant FEMH-2022-C-058); and Taipei Veterans General Hospital (grants V111E-002 and V111E-005-2). The sponsors had no role in the study design, data collection and analysis, publication decision, or manuscript drafting.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Counts of ICD-10-CM labels from the three hospitals. (A) Ranking of counts of labels in a medical record. (B) Ranking of counts of ICD-10-CM codes. ICD-10-CM: *International Classification of Diseases, 10th Revision, Clinical Modification*.
[DOCX File , 662 KB - medinform_v10i11e41342_app1.docx]

Multimedia Appendix 2
The punctuation count and the top 10 Chinese characters.
[DOCX File , 17 KB - medinform_v10i11e41342_app2.docx]

References

1. World Health Organization. International Statistical Classification of Diseases and Related Health Problems, 10th Revision: Volume 1, Tabular List, Fifth Edition 2016. Geneva, Switzerland: World Health Organization; 2016.
2. Mills RE, Butler RR, McCullough EC, Bao MZ, Averill RF. Impact of the transition to ICD-10 on Medicare inpatient hospital payments. Medicare Medicaid Res Rev 2011 Jun 06;1(2):001.02.a02 [FREE Full text] [doi: 10.5600/mmrr.001.02.a02] [Medline: 22340773]
3. Kusnoor SV, Blasingame MN, Williams AM, DesAutels SJ, Su J, Giuse NB. A narrative review of the impact of the transition to ICD-10 and ICD-10-CM/PCS. JAMIA Open 2019 Dec 26;3(1):126-131 [FREE Full text] [doi: 10.1093/jamiaopen/ooz066] [Medline: 32607494]
4. Chen PF, Wang SM, Liao WC, Kuo LC, Chen KC, Lin YC, et al. Automatic ICD-10 coding and training system: Deep neural network based on supervised learning. JMIR Med Inform 2021 Aug 31;9(8):e23230 [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
5. Shi H, Xie P, Hu Z, Zhang M, Xing EP. Towards automated ICD coding using deep learning. arXiv. Preprint posted online on November 11, 2017 [FREE Full text]
6. Sammani A, Bagheri A, van der Heijden PGM, Te Riele ASJM, Baas AF, Oosters CAJ, et al. Automatic multilabel detection of ICD10 codes in Dutch cardiology discharge letters using neural networks. NPJ Digit Med 2021 Feb 26;4(1):37 [FREE Full text] [doi: 10.1038/s41746-021-00404-9] [Medline: 33637859]
7. Wang X, Han J, Li B, Pan X, Xu H. Automatic ICD-10 coding based on multi-head attention mechanism and gated residual network. 2022 Presented at: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 9-12, 2021; Houston, TX p. 536-543. [doi: 10.1109/bibm52615.2021.9669625]
8. Makohon I, Li Y. Multi-label classification of ICD-10 coding and clinical notes using MIMIC and CodiEsp. 2021 Presented at: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI); July 27-30, 2021; Athens, Greece p. 1-4. [doi: 10.1109/bhi50953.2021.9508541]
9. Silva S, Gutman BA, Romero E, Thompson PM, Altmann A, Lorenzi M. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. 2019 Presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); April 8-11, 2019; Venice, Italy p. 270-274. [doi: 10.1109/isbi.2019.8759317]
10. Gao D, Ju C, Wei X, Liu Y, Chen T, Yang Q. HHHFL: Hierarchical heterogeneous horizontal federated learning for electroencephalography. arXiv. Preprint posted online on September 11, 2019 [FREE Full text]
11. Liu D, Dligach D, Miller T. Two-stage federated phenotyping and patient representation learning. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August 1, 2019; Florence, Italy p. 283-291 URL: https://aclanthology.org/W19-5030v1.pdf [doi: 10.18653/v1/w19-5030]
12. Wang SM, Chang YH, Kuo LC, Lai F, Chen YN, Yu FY, et al. Using deep learning for automatic Icd-10 classification from free-text data. Eur J Biomed Inform (Praha) 2020;16(1):1-10 [FREE Full text] [doi: 10.24105/ejbi.2020.16.1.1]
13. Chen PF, Chen KC, Liao WC, Lai F, He TL, Lin SC, et al. Automatic International Classification of Diseases coding system: Deep contextualized language model with rule-based approaches. JMIR Med Inform 2022 Jun 29;10(6):e37557 [FREE Full text] [doi: 10.2196/37557] [Medline: 35767353]
14. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2022 Jan;3(1):1-23. [doi: 10.1145/3458754]
15. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 29, 2019 [FREE Full text]
16. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on April 10, 2019 [FREE Full text]
17. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
18. McMahan HB, Moore E, Ramage D, y Arcas BA. Federated learning of deep networks using model averaging. arXiv. Preprint posted online on Februrary 17, 2016 [FREE Full text]
19. Beutel DJ, Topal T, Mathur A, Qiu X, Parcollet T, Lane ND. Flower: A friendly federated learning research framework. arXiv. Preprint posted online on July 28, 2020 [FREE Full text]
20. gRPC Authors. gRPC: A high performance, open source universal RPC framework. gRPC. URL: https://grpc.io [accessed 2022-09-17]
21. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable prediction of medical codes from clinical text. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, Louisiana p. 1101-1111 URL: https://aclanthology.org/N18-1100.pdf [doi: 10.18653/v1/n18-1100]

22. Li L, Fan Y, Tse M, Lin KY. A review of applications in federated learning. Comput Ind Eng 2020 Nov;149:106854. [doi: 10.1016/j.cie.2020.106854]
23. Imteaj A, Thakker U, Wang S, Li J, Amini MH. A survey on federated learning for resource-constrained IoT devices. IEEE Internet Things J 2022 Jan 1;9(1):1-24. [doi: 10.1109/jiot.2021.3095077]
24. Zhao S, Gupta R, Song Y, Zhou D. Extremely small BERT models from mixed-vocabulary training. 2021 Presented at: 16th Conference of the European Chapter of the Association for Computational Linguistics; April 19-23, 2021; Online p. 2753-2759 URL: https://aclanthology.org/2021.eacl-main.238.pdf [doi: 10.18653/v1/2021.eacl-main.238]
25. Biswas B, Pham TH, Zhang P. TransICD: Transformer based code-wise attention model for explainable ICD coding. 2021 Presented at: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021; June 15-18, 2021; Virtual Event p. 469-478. [doi: 10.1007/978-3-030-77211-6_56]
26. Abdullah M, AlMasawa M, Makki I, Alsolmi M, Mahrous S. Emotions extraction from Arabic tweets. International Journal of Computers and Applications 2018 Jun 07;42(7):661-675. [doi: 10.1080/1206212x.2018.1482395]
27. Ek A, Bernardy JP, Chatzikyriakidis S. How does punctuation affect neural models in natural language inference. 2020 Presented at: Probability and Meaning Conference (PaM 2020); October 14-15, 2020; Gothenburg, Sweden p. 109-116 URL: https://aclanthology.org/2020.pam-1.15.pdf

Abbreviations

BERT: bidirectional encoder representations from transformers
BioBERT: Bidirectional Encoder Representations From Transformers for Biomedical Text Mining
CC-News: CommonCrawl News
FEMH: Far Eastern Memorial Hospital
gRPC: Google Remote Procedure Call
ICD-10: *International Classification of Diseases, 10th Revision*
ICD-10-CM: *International Classification of Diseases, 10th Revision, Clinical Modification*
ICD: *International Classification of Diseases*
MIMIC-III: Medical Information Mart for Intensive Care III
NLP: natural language processing
NTUH: National Taiwan University Hospital
RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach
VGHTPE: Taipei Veterans General Hospital

Edited by C Lovis; submitted 24.07.22; peer-reviewed by I Li, N Nuntachit; comments to author 15.08.22; revised version received 03.10.22; accepted 08.10.22; published 10.11.22.

Please cite as:
Chen PF, He TL, Lin SC, Chu YC, Kuo CT, Lai F, Wang SM, Zhu WX, Chen KC, Kuo LC, Hung FM, Lin YC, Tsai IC, Chiu CH, Chang SC, Yang CY
Training a Deep Contextualized Language Model for International Classification of Diseases, 10th Revision Classification via Federated Learning: Model Development and Validation Study
JMIR Med Inform 2022;10(11):e41342
URL: https://medinform.jmir.org/2022/11/e41342
doi:10.2196/41342
PMID:36355417

Original Paper

Automatic Screening of Pediatric Renal Ultrasound Abnormalities: Deep Learning and Transfer Learning Approach

Ming-Chin Tsai[1], MD; Henry Horng-Shing Lu[2], PhD; Yueh-Chuan Chang[3], MSc; Yung-Chieh Huang[1,4], MD; Lin-Shien Fu[1,4,5], MD

[1]Department of Pediatrics, Taichung Veterans General Hospital, Taichung, Taiwan

[2]Institute of Statistics, National Yang Ming Chiao Tung University, Hsing-chu, Taiwan

[3]Institute of Electrical & Control Engineering, National Yang Ming Chiao Tung University, Hsing-chu, Taiwan

[4]Department of Pediatrics, National Yang Ming Chiao Tung University, Taipei, Taiwan

[5]Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung, Taiwan

Corresponding Author:
Lin-Shien Fu, MD
Department of Pediatrics
Taichung Veterans General Hospital
No.1650, Section 4, Taiwan Blvd.
Taichung
Taiwan
Phone: 886 4 23592525 ext 5909
Fax: 886 4 23741359
Email: linshienfu@yahoo.com.tw

Abstract

Background: In recent years, the progress and generalization surrounding portable ultrasonic probes has made ultrasound (US) a useful tool for physicians when making a diagnosis. With the advent of machine learning and deep learning, the development of a computer-aided diagnostic system for screening renal US abnormalities can assist general practitioners in the early detection of pediatric kidney diseases.

Objective: In this paper, we sought to evaluate the diagnostic performance of deep learning techniques to classify kidney images as normal and abnormal.

Methods: We chose 330 normal and 1269 abnormal pediatric renal US images for establishing a model for artificial intelligence. The abnormal images involved stones, cysts, hyperechogenicity, space-occupying lesions, and hydronephrosis. We performed preprocessing of the original images for subsequent deep learning. We redefined the final connecting layers for classification of the extracted features as abnormal or normal from the ResNet-50 pretrained model. The performances of the model were tested by a validation data set using area under the receiver operating characteristic curve, accuracy, specificity, and sensitivity.

Results: The deep learning model, 94 MB parameters in size, based on ResNet-50, was built for classifying normal and abnormal images. The accuracy, (%)/area under curve, of the validated images of stone, cyst, hyperechogenicity, space-occupying lesions, and hydronephrosis were 93.2/0.973, 91.6/0.940, 89.9/0.940, 91.3/0.934, and 94.1/0.996, respectively. The accuracy of normal image classification in the validation data set was 90.1%. Overall accuracy of (%)/area under curve was 92.9/0.959..

Conclusions: We established a useful, computer-aided model for automatic classification of pediatric renal US images in terms of normal and abnormal categories.

(JMIR Med Inform 2022;10(11):e40878) doi:10.2196/40878

KEYWORDS

transfer learning; convolutional neural networks; pediatric renal ultrasound image; screening; pediatric; medical image; clinical informatics; deep learning; ultrasound image; artificial intelligence; diagnostic system

Introduction

Renal abnormalities are important findings in pediatric medicine. It is well accepted that "silent" renal abnormalities can be effectively detected through ultrasound (US) screening, which makes both early diagnoses and intervention possible [1,2]. US is a safe, relatively cheap, and convenient medical modality. Portable ultrasonic probes and internet connections have largely developed in recent years, even extending the coverage of pediatric renal US screening throughout the world. However, current methods remain limited due to the lack of automated processes that accurately classify diseased and normal kidneys [3].

Common renal abnormalities identified in US images in a series of more than 1 million school children included hydronephrosis (39.6%), unilateral small kidney (19.8%), unilateral agenesis (15.9%), cystic disease (13.9%), abnormal shapes—ectopic, horseshoe, and duplication of kidney (8%)—as well as others, that is, stones, tumors, and parenchymal diseases (1.5%) [1].

Thus far, publications regarding computer-aided US image interpretation have been much fewer than those based on computerized tomography or magnetic resonance imaging [4,5]. The use of US presents unique challenges, such as different angles of image sampling, low image quality caused by noise and artifacts, high dependence on an abundance of operators, and high inter- and intra-observer variability across different institutes and manufacturers' US systems [6]. From the review about medical US published in 2021 [7], there were only 3 studies involving deep learning for renal US image classification [5,8,9].

This study was performed to select normal pediatric renal US images, as well as different types of renal abnormalities previously mentioned, for purposes of machine learning. Through the pretreatment of original images, adequate grouping of images, and deep neural network training, we hope that renal images can be correctly classified as either normal or abnormal. The aim of this study is to establish an artificial intelligence (AI) model for screening renal abnormalities to enhance the well-being of children even in areas where there is no pediatric nephrologist.

Methods

Ethics Approval

This study was approved by the institutional review board of Taichung Veterans General Hospital (No. CE20204A).

Materials

The images used were all created from the original images in the pediatric US examination room at Taichung Veterans General Hospital from January 2000 to December 2020. Here were 4 different US machines manufactured by both Philips and Acuson, which were used in this study. All images were obtained by a US technician having more than 20 years of experience, using a 4 MHz sector transducer. We chose only images taken of a longitudinal view from the right and left kidney.

We established 2 data sets. One data set was for training, and the other was for validation. The images in these 2 data sets were totally different.

Image Preprocessing and Data Cleaning

All images were detached from their original general data, including name, date of birth, date of examination, and chart number. The size of all the images was 600x480 pixels. We processed the images using software to obtain adequate illustrations for machine learning. As shown in Figure 1, after preprocessing, the images contain a kidney, a square of liver obtained from the examination simultaneously, and a gray scale gradient seen in the left upper part of the image.

Figure 1. Preprocessing images for machine learning.

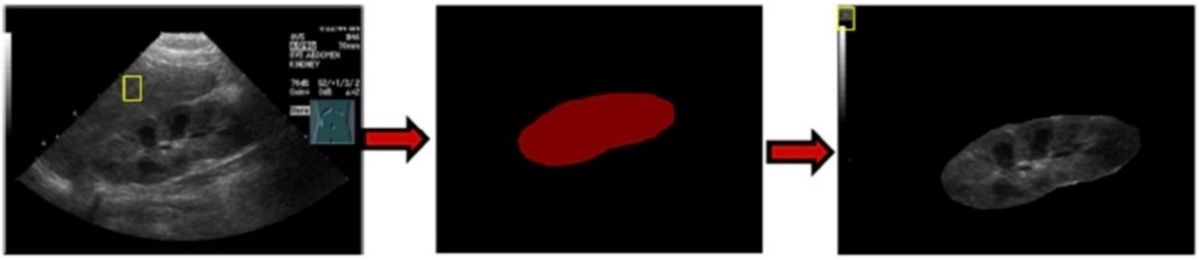


Image Grouping

Normal images were those having a normal size and shape, as well as a clear renal cortex or medulla without hydronephrosis, hyperechogenicity, cysts, stones, or any space-occupying lesion. We prepared 330 images for this group. There were a total of 1269 abnormal renal images. The abnormalities included hydronephrosis, hyperechogenicity, cysts, stones, and space-occupying lesions. The number of images and examinations are summarized in Table 1. The hyperechogenicity of the renal US images included increased renal cortex echogenicity as compared to the liver, a poor differentiation of the renal cortex or medulla, and an inversed echogenicity of the renal cortex or medulla. These findings were judged by 2 pediatric nephrologists.

Table 1. Distribution of images and examinations in the training and testing augmented database.

Diagnosis	Training (cases/images)	Testing (cases/images)	Totals (cases/images)
Normal	132/264	32/66	164/330
Abnormal			
Stone	146/342	37/85	183/427
Cyst	100/215	25/53	125/268
Hyperechogenicity	60/132	15/33	75/165
Space-occupying lesions	108/181	26/45	134/226
Hydronephrosis	68/146	16/37	84/183
Total	614/1280	151/319	765/1599

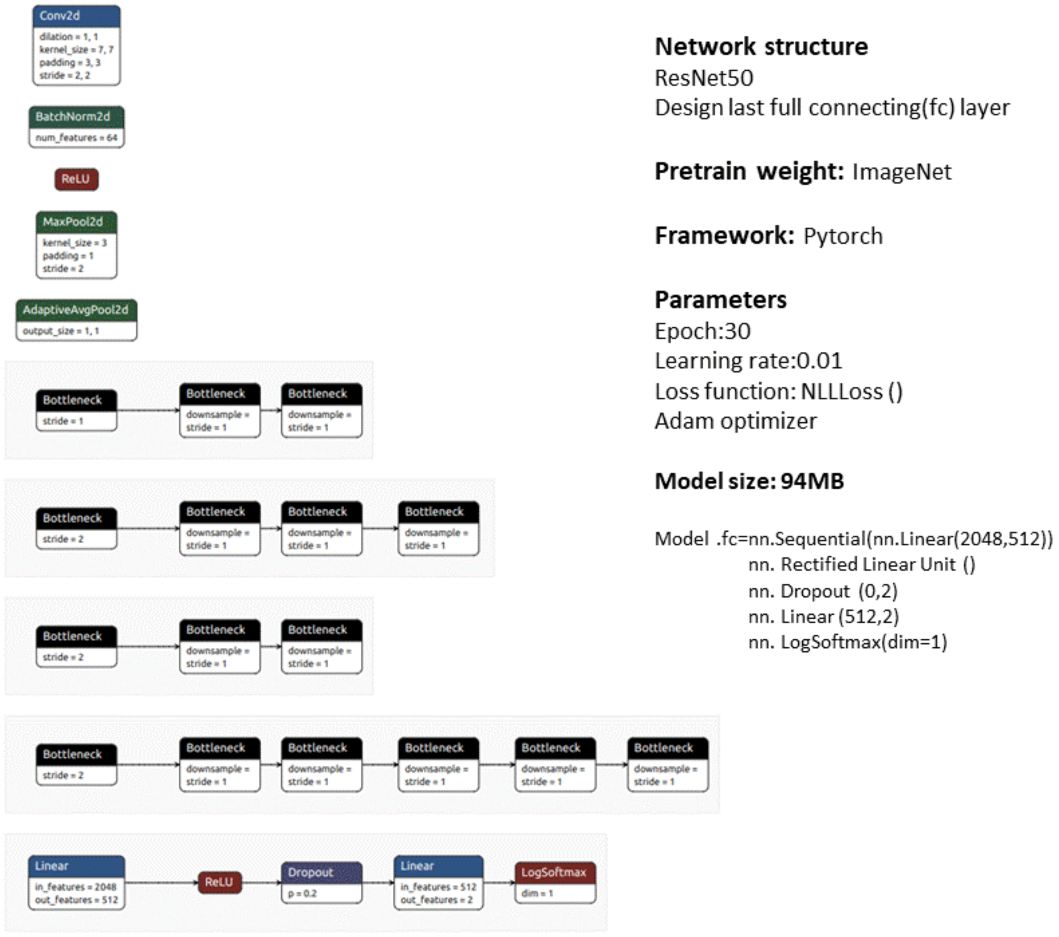
Machine Learning

We performed feature extraction with the pretrained model of ResNet-50 [8-10] in PyTorch from the data set ImageNet [11]. We used the pretrained weight of ResNet, so there was no backpropagation during feature extraction for training US images. The input data used were renal US images of 800x600 pixels in size. We normalized the dimension to 224x224 pixels prior to feeding the images into the network.

For the classification purpose, we redefined the final fully connected layers, which output image classification as abnormal or normal. After the training images went through Resnet50, there were 2048 outputs. There were 4 components in the final fully connected layer. The first was a linear layer with the 2048 feature extractions and 512 outputs. The second was rectified linear unit, which was a piecewise linear function that only outputted the positive result. Subsequently, we added the dropout layer to prevent overfitting. The 4th component was another linear layer, performing with 512 inputs and 2 outputs, which stand for the 2 categories, that is, abnormal and normal class with their probabilities.

We optimized the model with the Adam optimizer at a learning rate of 0.01 [12]. There were a total of 30 epochs used for convolutional neural network training. We created a 94 MB size model to classify normal versus abnormal renal US images. Figure 2 is a summary of our deep learning structure.

Figure 2. Brief summary of machine learning.



Network structure
ResNet50
Design last full connecting(fc) layer

Pretrain weight: ImageNet

Framework: Pytorch

Parameters
Epoch:30
Learning rate:0.01
Loss function: NLLLoss ()
Adam optimizer

Model size: 94MB

```
Model .fc=nn.Sequential(nn.Linear(2048,512))
          nn. Rectified Linear Unit ()
          nn. Dropout (0,2)
          nn. Linear (512,2)
          nn. LogSoftmax(dim=1)
```

Experimental Setup

We implemented the training-testing approach. The data set was randomly divided into 1272/1599 (79.55%) images for training and 327/1599 (20.45%)images for testing to establish the model. We performed a 10-time randomization of the data set to repeat the machine learning described in the previous paragraph. For validation of the 94 MB model, there was another validation data set with 327 pediatric renal US images, including 66 (20.2%) normal, 37 (11.3%) hydronephrosis, 53 (16.2%) cyst, 95 (29.1%) stone, 53 (16.2%) hyperechogenicity, and 26 (7.9%) space-occupying US images. All these images were totally different from the data set for establishing the model.

Evaluation of Performance

We evaluated the performance from a single image result. The diagnostic performance was measured by accuracy, specificity, sensitivity, positive predictive value, and negative predictive value. To calculate the above metrics, we defined an abnormal result as positive and a normal result as negative.

Results

After 30 epochs for these 1599 pediatric renal US images, we obtained satisfactory results. The performance metrics in the test part of the data set are shown in Table 2. The accuracy in different abnormalities ranged from 95% to 100%.

Table 2. Evaluation metrics for screening different abnormalities from test renal ultrasound images in the data set.

Diagnosis (number)	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC[a]	PPV[b] (%)	NPV[c] (%)
Stone	100	100	100	0.974	100	100
Cyst	95.2	88.5	100	0.945	100	91.7
Hyperechogenicity	98.3	96.2	100	0.938	100	97.1
Space-occupying lesions	98.7	95.6	100	0.935	100	97.1
Hydronephrosis	100	100	100	0.998	100	100
Overall	98.4	96.39	100	0.961	100	97.2

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

[b] PPV: positive predictive value.

[c] NPV: negative predictive value.

The accuracies of each abnormality ranged from 95.2% to 100%, with an overall accuracy as 98.4%. The area under curves (AUCs) were from 0.935 to 0.998. The AUC for overall performance was 0.961. There was no difference between these 10 random tests (*P*>.05). We repeated the 10 experiments using different randomizations involving 80%/20% training/test images to check the consistency of the machine learning performance. The accuracies ranged from 95.2% to 98.4%. There was no difference between these 10 tests (*P*>.05). We performed a 5-fold cross test, and the results are shown in Table 3.

We validated the 94 MB model through machine learning with another 327 pediatric renal US images. The classifications included 66 (20.2%) normal, 37 (11.3%) hydronephrosis, 53 (16.2%) cyst, 95 (29.1%) stone, 53 (16.2%) hyperechogenicity, and 26 (7.9%) space-occupying US images. The performances based on each single image are summarized in Table 4. Accuracy in the different abnormalities ranged from 89.9% to 94.1%, with an average of 92.3%. AUC was from 0.934 to 0.996 (Figure 3). The overall performance in AUC was 0.959. The macro F_1 was 0.924.

Table 3. Results of the 5-fold cross test.

	Test 1	Test 2	Test 3	Test 4	Test 5	Overall
Normal accuracy (%)	80	87.9	87.9	87.9	87.9	86.32
Stone accuracy (%)/AUC[a]	91.2/0.925	92.9/0.897	89.4/0.923	89.4/0.925	94.3/0.927	91.60/0.927
Cyst accuracy (%)/AUC	75.4/0.858	90.6/0.896	84.9/0.927	90.6/0.898	82.1/0.891	85.3/0.903
hyperechogenicity accuracy (%)/AUC	84.8/0.848	81.8/0.855	81.8/0.862	81.8/0.862	81.8/0.891	84.2/0.859
Space-occupying lesion accuracy (%)/AUC	92.5/0.903	84.9/0.881	94.5/0.917	83.0/0.874	82.6/0.863	86.8/0.896
Hydronephrosis accuracy (%)/AUC	100/0.965	91.9/0.888	89.2/0.940	94.6/0.932	91.4/0.871	94/0.928
Overall accuracy (%)/AUC	87.8/0.903	89/0.887	87.8/0.928	87.5/0.902	87.7/0.901	88.3/0.900

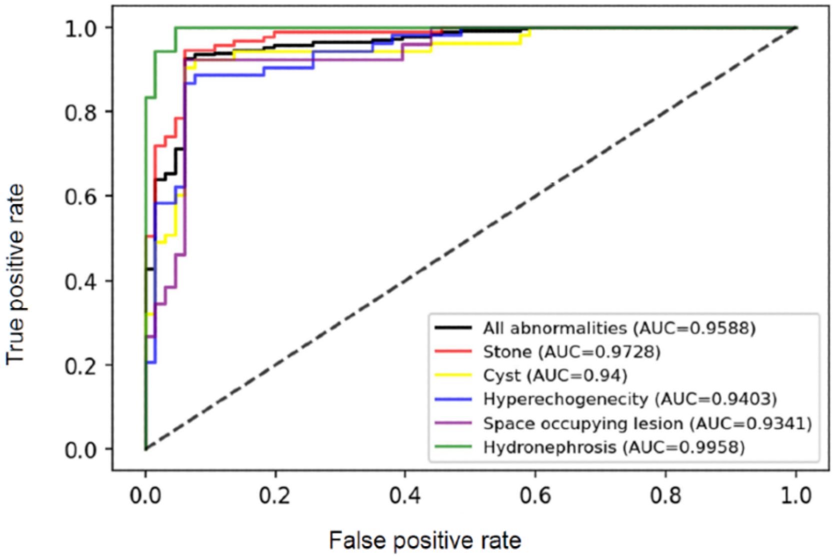
[a] AUC: area under curve.

Table 4. Evaluation metrics for screening different abnormalities from other renal ultrasound images for validation.

Diagnosis	US images, n (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC[a]	PPV[b] (%)	NPV[c] (%)	F_1-score
Normal	66 (20.2)	N/A[d]	N/A	90.9%	N/A	N/A	N/A	N/A
Stone	93 (28.4)	93.2	94.7	N/A	0.973	93.2	92.3	0.927
Cyst	53 (16.2)	91.6	92.5	N/A	0.940	91.6	93.8	0.918
Hyperechogenicity	53 (16.2)	89.9	88.7	N/A	0.940	89.9	90.9	0.897
Space-occupying lesions	26 (7.9)	91.3	92.3	N/A	0.934	91.3	96.81	0.923
Hydronephrosis	37 (11.3)	94.1	100	N/A	0.996	94.2	100	0.957
Overall	328 (100)	92.9	96.1	N/A	0.959	93.6	77.92	0.924[e]

[a] AUC-ROC: area under the receiver operating characteristic curve.
[b] PPV: positive predictive value.
[c] NPV: negative predictive value.
[d] N/A: not applicable.
[e] Macro F_1.

Figure 3. Area under the receiver operating characteristic curves of different image abnormalities and the overall performance. AUC: area under curve.



Discussion

The main finding of this study is a useful AI model for screening abnormal pediatric renal US images. The average accuracy can be 92.9%. The results can fulfill the main purpose of this study—to develop a useful computer-aided diagnosis model for screening various pediatric renal US abnormal patterns automatically. In this study, the machine learning methods were based upon convolutional neural network and fine-tuning, along with our unique methods for image preprocessing, as well as strategies for classification, which achieved a feasible model for clinical purposes. We constructed the stable classifier that combined both the transfer learning and training from scratch, balancing the training of a medical data set taken from an adequate sample size.

Clinical applications of AI in nephrology are versatile, but the use of renal US in this field is still in its infancy [13,14]. The reports derived from renal US images alone have been relatively limited up until now, with the major reports involving acute and chronic injuries [15-17]. Most renal image studies for AI used magnetic resonance imaging, computerized tomography, and patient histology for tumors, stones, nephropathy, transplantation, and other conditions [18-21]. The key challenges associated with deep learning involving US include reliability, generalizability, and bias [22]. The basic studies for enhancing AI performance in renal US have begun and remain undergoing [23-25].

There have been 4 reports from studies involving clinical AI applications in pediatric renal US abnormalities [3, 5,8,9]. Zheng et al [3] found that the deep transfer learning method offers satisfactory accuracy in identifying congenital anomalies in the kidney and urinary tract, even when the data set is as small as only having 50 children with congenital anomalies in the kidney and urinary tract and 50 children as the control. Yin et al [5] performed a similar study to detect posterior urethral valves. Sudarharson et al [8] used 3 variant data sets for identifying renal cysts, stones, and tumors, with an accuracy rate of 96.54% in images of quality and 95.58% in images of noise. Smail et al [9] attempted to use AI for grading hydronephrosis involving the 5-point scoring system from the Society of Fetal Urology (SFU). The best recorded performance was a 78% accuracy rate by dividing hydronephrosis into mild and severe. However, the accuracy rate was only 51% when using the 5-point system. In our study, we established a single 94 MB model to classify normal versus abnormal pediatric renal US images. The items seen in the abnormalities included renal cysts, stones, and tumors, as reported by Sudarharson et al [8]. In addition, the model was able to identify images of hydronephrosis and hyperechogenicity. Comparing the results from the study performed by Smail et al [9], our results showed a better classification accuracy for hydronephrosis. The 37 validated images were moderate or severe hydronephrosis, that is, the SFU class II, III, and IV. Our model can achieve 100% sensitivity, comparing the sensitivity of 46%-54%, as previously reported [26].

In terms of SFU class I, our model had an accuracy of 71.7% (119/166). Up until now, grading of hydronephrosis has been an ongoing challenge [27]. Extremely early intervention for treatment of mild hydronephrosis remains inadequate. If a child with mild hydronephrosis is also experiencing other renal abnormalities, such as stones, cysts, or hyperechogenicity, it is highly possible our model would be capable of providing any alarming information surrounding these conditions.

The unique pretreatment of images for machine learning performed in this study was performed to provide a comparison of liver echogenicity in the simultaneous study. This step is necessary for identifying hyperechogenicity. Other abnormalities, such as hydronephrosis, cysts, stones, and tumors, showed no difference in classification, regardless of whether we inputted the images with the addition of the square containing liver echogenicity and the gray scale gradient in the left part of the image shown in Figure 1. As demonstrated in Table 4, the accuracy and sensitivity for hyperechogenicity identification was lower than it was with other abnormalities. Increased echogenicity is an important finding in evaluating muscle, thyroid, vascular, and renal diseases [28]. The gray scale US presents a general sensitivity rate of 62% to 77%, a specificity of 58% to 73%, and a positive predictive value of 92% for detecting microscopically confirmed renal parenchymal diseases. The above results reveal that the echogenicity change was not sensitive enough for detecting renal disease. Abnormalities in renal echogenicity include increased echogenicity, poor differentiation of the cortex or medulla, and inversed echogenicity of the renal cortex and medulla [29]. In practice, it is quite often that we cannot obtain a square containing homogenous liver echogenicity for purposes of machine learning. When the classification is compared by a pediatric nephrologist, the results are acceptable. It is also difficult for the naked eye to discriminate between the not-so-significant gray scale differences. Currently, the so called "radiomics" information, which can aid US imaging in AI, is emerging [30], with a more precise assessment of US pixels possibly enhancing the utility of hyperechogenicity.

A limitation of this study is the single medical center image source. More images from different hospitals, areas, ethnicities, and US companies need to be used. We conducted a small-scale external validation using US images from different companies, including General Electric, Siemens, and Toshiba. After image pretreatment, the results could be 100% sensitivity, 80% specificity, and 90% accuracy. Another limitation is the moderate image number of images contributing to the data set. We did not divide images from right or left kidney for training, though the results can be acceptable. We will further validate our method based on larger data sets.

In conclusion, this study proposed the use of an automatic model for purposes of screening various abnormalities in pediatric renal US images. We will continue to enhance the model's performance as we conduct additional evaluation studies surrounding its future clinical applications, including being an auxiliary software for screening children's renal abnormalities in remote areas.

Acknowledgments

This study was supported in part by a grant from Taichung Veterans General Hospital (No. TCVGH-1106506B, TCVGH-1116504C). The statistical work is partially supported by the Ministry of Science and Technology, Taiwan, R.O.C. (grant MOST 110-2118-M-A49-002-MY3 and 110-2634-F-A49-005).

Conflicts of Interest

None declared.

References

1. Sheih CP, Liu MB, Hung CS, Yang KH, Chen WY, Lin CY. Renal abnormalities in schoolchildren. Pediatrics 1989 Dec;84(6):1086-1090. [Medline: 2685739]
2. Parakh P, Bhatta NK, Mishra OP, Shrestha P, Budhathoki S, Majhi S, et al. Urinary screening for detection of renal abnormalities in asymptomatic school children. Nephrourol Mon 2012;4(3):551-555 [FREE Full text] [doi: 10.5812/numonthly.3528] [Medline: 23573484]
3. Zheng Q, Furth SL, Tasian GE, Fan Y. Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features. J Pediatr Urol 2019 Feb;15(1):75.e1-75.e7 [FREE Full text] [doi: 10.1016/j.jpurol.2018.10.020] [Medline: 30473474]

4. Akkus Z, Cai J, Boonrod A, Zeinoddini A, Weston AD, Philbrick KA, et al. A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow. J Am Coll Radiol 2019 Sep;16(9 Pt B):1318-1328. [doi: 10.1016/j.jacr.2019.06.004] [Medline: 31492410]
5. Yin S, Peng Q, Li H, Zhang Z, You X, Fischer K, et al. Multi-instance Deep Learning of Ultrasound Imaging Data for Pattern Classification of Congenital Abnormalities of the Kidney and Urinary Tract in Children. Urology 2020 Aug;142:183-189 [FREE Full text] [doi: 10.1016/j.urology.2020.05.019] [Medline: 32445770]
6. Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, et al. Deep Learning in Medical Ultrasound Analysis: A Review. Engineering 2019 Apr;5(2):261-275. [doi: 10.1016/j.eng.2018.11.020]
7. De Jesus-Rodriguez HJ, Morgan MA, Sagreiya H. Deep Learning in Kidney Ultrasound: Overview, Frontiers, and Challenges. Adv Chronic Kidney Dis 2021 May;28(3):262-269. [doi: 10.1053/j.ackd.2021.07.004] [Medline: 34906311]
8. Sudharson S, Kokil P. An ensemble of deep neural networks for kidney ultrasound image classification. Comput Methods Programs Biomed 2020 Dec;197:105709. [doi: 10.1016/j.cmpb.2020.105709] [Medline: 32889406]
9. Smail LC, Dhindsa K, Braga LH, Becker S, Sonnadara RR. Using Deep Learning Algorithms to Grade Hydronephrosis Severity: Toward a Clinical Adjunct. Front Pediatr 2020;8:1 [FREE Full text] [doi: 10.3389/fped.2020.00001] [Medline: 32064241]
10. ResNet. PyTorch. URL: https://pytorch.org/hub/pytorch_vision_resnet [accessed 2022-10-05]
11. Fan R, Chang K, Hsieh C, Wang XR, Lin CJ. LIBLINEAR: a library for large linear classification. JMLR 2008;9(61):1871-1874.
12. He K, Zhang X, Ren S. Deep residual learning for image recognition. 2016 Presented at: Proceedings of the IEEE conference on computer vision and pattern recognition; June 27-30, 2016; Las Vegas, NV, USA. [doi: 10.1109/cvpr.2016.90]
13. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng 2010 Oct;22(10):1345-1359. [doi: 10.1109/tkde.2009.191]
14. Abadi M. TensorFlow: learning functions at scale. SIGPLAN Not 2016 Dec 05;51(9):1-1. [doi: 10.1145/3022670.2976746]
15. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. JMLR 2014;15(56):1929-1958.
16. Rashidi P, Bihorac A. Artificial intelligence approaches to improve kidney care. Nat Rev Nephrol 2020 Feb;16(2):71-72 [FREE Full text] [doi: 10.1038/s41581-019-0243-3] [Medline: 31873197]
17. Lee M, Wei S, Anaokar J, Uzzo R, Kutikov A. Kidney cancer management 3.0: can artificial intelligence make us better? Curr Opin Urol 2021 Jul 01;31(4):409-415. [doi: 10.1097/MOU.0000000000000881] [Medline: 33882560]
18. Kuo C, Chang C, Liu K, Lin W, Chiang H, Chung C, et al. Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning. NPJ Digit Med 2019;2:29 [FREE Full text] [doi: 10.1038/s41746-019-0104-2] [Medline: 31304376]
19. Bandara MS, Gurunayaka B, Lakraj G, Pallewatte A, Siribaddana S, Wansapura J. Ultrasound Based Radiomics Features of Chronic Kidney Disease. Acad Radiol 2022 Feb;29(2):229-235. [doi: 10.1016/j.acra.2021.01.006] [Medline: 33589307]
20. Ying F, Chen S, Pan G, He Z. Artificial Intelligence Pulse Coupled Neural Network Algorithm in the Diagnosis and Treatment of Severe Sepsis Complicated with Acute Kidney Injury under Ultrasound Image. J Healthc Eng 2021;2021:6761364 [FREE Full text] [doi: 10.1155/2021/6761364] [Medline: 34336164]
21. Nikpanah M, Xu Z, Jin D, Farhadi F, Saboury B, Ball MW, et al. A deep-learning based artificial intelligence (AI) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic MRI. Clin Imaging 2021 Sep;77:291-298. [doi: 10.1016/j.clinimag.2021.06.016] [Medline: 34171743]
22. Yildirim K, Bozdag PG, Talo M, Yildirim O, Karabatak M, Acharya UR. Deep learning model for automated kidney stone detection using coronal CT images. Comput Biol Med 2021 Aug;135:104569. [doi: 10.1016/j.compbiomed.2021.104569] [Medline: 34157470]
23. Hermsen M, Volk V, Bräsen JH, Geijs DJ, Gwinner W, Kers J, et al. Quantitative assessment of inflammatory infiltrates in kidney transplant biopsies using multiplex tyramide signal amplification and deep learning. Lab Invest 2021 Aug;101(8):970-982 [FREE Full text] [doi: 10.1038/s41374-021-00601-w] [Medline: 34006891]
24. Farris AB, Vizcarra J, Amgad M, Cooper LAD, Gutman D, Hogan J. Artificial intelligence and algorithmic computational pathology: an introduction with renal allograft examples. Histopathology 2021 May;78(6):791-804 [FREE Full text] [doi: 10.1111/his.14304] [Medline: 33211332]
25. De Jesus-Rodriguez HJ, Morgan MA, Sagreiya H. Deep Learning in Kidney Ultrasound: Overview, Frontiers, and Challenges. Adv Chronic Kidney Dis 2021 May;28(3):262-269. [doi: 10.1053/j.ackd.2021.07.004] [Medline: 34906311]
26. Chen G, Dai Y, Zhang J, Yin X, Cui L. MBANet: Multi-branch aware network for kidney ultrasound images segmentation. Comput Biol Med 2022 Feb;141:105140. [doi: 10.1016/j.compbiomed.2021.105140] [Medline: 34922172]
27. Lassau N, Estienne T, de Vomecourt P, Azoulay M, Cagnol J, Garcia G, et al. Five simultaneous artificial intelligence data challenges on ultrasound, CT, and MRI. Diagn Interv Imaging 2019 Apr;100(4):199-209. [doi: 10.1016/j.diii.2019.02.001] [Medline: 30885592]
28. Onen A. Grading of Hydronephrosis: An Ongoing Challenge. Front Pediatr 2020;8:458 [FREE Full text] [doi: 10.3389/fped.2020.00458] [Medline: 32984198]

29. Quaia E, Correas JM, Mehta M, Murchison JT, Gennari AG, van Beek EJR. Gray Scale Ultrasound, Color Doppler Ultrasound, and Contrast-Enhanced Ultrasound in Renal Parenchymal Diseases. Ultrasound Q 2018 Dec;34(4):250-267. [doi: 10.1097/RUQ.0000000000000383] [Medline: 30169495]
30. Grenier N, Merville P, Combe C. Radiologic imaging of the renal parenchyma structure and function. Nat Rev Nephrol 2016 Jun;12(6):348-359. [doi: 10.1038/nrneph.2016.44] [Medline: 27067530]

Abbreviations

AI: artificial intelligence
AUC: area under curve
SFU: Society of Fetal Urology
US: ultrasound

Edited by C Lovis; submitted 12.07.22; peer-reviewed by M Lee, Y Fan, SJC Soerensen , V Khetan, S Yin; comments to author 03.08.22; revised version received 16.09.22; accepted 02.10.22; published 02.11.22.

Please cite as:
Tsai MC, Lu HHS, Chang YC, Huang YC, Fu LS
Automatic Screening of Pediatric Renal Ultrasound Abnormalities: Deep Learning and Transfer Learning Approach
JMIR Med Inform 2022;10(11):e40878
URL: https://medinform.jmir.org/2022/11/e40878
doi:10.2196/40878
PMID:36322109

Corrigenda and Addenda

Correction: 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:

Correction of: https://medinform.jmir.org/2022/5/e33219

(JMIR Med Inform 2022;10(11):e43520) doi:10.2196/43520

In “Web-Based Software Tools for Systematic Literature Review in Medicine: Systematic Search and Feature Analysis” (JMIR Med Inform 2022;10(5):e33219) the authors noted some errors and made the following corrections:

1. For the “Access” category in Table 4, features included free, living, public outputs, and multiple users. In the originally published article, the feature "public outputs" was not counted, understating the total features offered. Therefore, Table 4 has been revised, as follows:

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)

The originally published Table 4 can be found in Multimedia Appendix 1.

Accordingly, the in-text references to Table 4 were revised in the article, as follows:

2. In the originally published article, in the Abstract, the section "Results" was the following:

> *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%. DistillerSR (Evidence Partners; 26/30, 87%), Nested Knowledge (Nested Knowledge; 25/30, 83%), and EPPI-Reviewer Web (EPPI-Centre; 24/30, 80%) support the most features followed by Giotto Compliance (Giotto Compliance; 23/30, 77%), LitStream (ICF), and SRDB.PRO (VTS Software). 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.*

In the Abstract, the section "Results" has been revised, as follows:

> *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.*

3. In the originally published article, under Methods, the first paragraph of the section “Evaluation of Tools” was the following:

> *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, 50% (4/8) did not respond, 38% (3/8) 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.*

The first paragraph of the section “Evaluation of Tools” has been revised, as follows:

> *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.*

4. In the originally published article, under Results, the section "Feature Assessment" was the following:

> *DistillerSR (26/30, 87%), Nested Knowledge (25/30, 83%), and EPPI-Reviewer Web (24/30, 80%) support the most features, followed by Giotto Compliance (23/30, 77%), LitStream, and SRDB.PRO (VTS Software). 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 5 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), RobotReviewer (RobotReviewer), and COVID-NMA (COVID-NMA; Table 3).*

The section “Feature Assessment” has been replaced, as follows:

> *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).*

5. In the originally published article, the section "Feature Assessment: Breakout by Feature Class" was the following:

> *Of all 6 feature classes, administrative features are the most supported, and extraction features are the least supported (Figure 3). Only 2 tools, Covidence (Cochrane) and EPPI-Reviewer, offer all 4 extraction features (Table 4). DistillerSR, Nested Knowledge, and JBI SUMARI (JBI) support all 4 documentation/output features.*

The section “Feature Assessment: Breakout by Feature Class” has been revised, as follows:

> *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.*

6. In the originally published article, under Discussion, the “Principal Findings” section was the following:

> *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 (DistillerSR, Nested Knowledge, EPPI-Reviewer, and Giotto Compliance) all offered between 80% and 87% 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.*

The section “Principal Findings” has been revised, as follows:

> *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.*

The authors confirm that these data changes do not affect the conclusions of the paper.

The correction will appear in the online version of the paper on the JMIR Publications website on November 23, 2022, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Original Table 4.
[PNG File , 1039 KB - medinform_v10i11e43520_app1.png]

Submitted 13.10.22; this is a non–peer-reviewed article;accepted 14.10.22; published 23.11.22.

Please cite as:
Cowie K, Rahmatullah A, Hardy N, Holub K, Kallmes K
Correction: Web-Based Software Tools for Systematic Literature Review in Medicine: Systematic Search and Feature Analysis
JMIR Med Inform 2022;10(11):e43520
URL: https://medinform.jmir.org/2022/11/e43520
doi:10.2196/43520
PMID:36417760

Review

Considering Clinician Competencies for the Implementation of Artificial Intelligence–Based Tools in Health Care: Findings From a Scoping Review

Kim V Garvey[1,2*], MLIS, PhD; Kelly Jean Thomas Craig[3,4*], PhD; Regina Russell[5], MEd, MA, PhD; Laurie L Novak[6,7], MHSA, PhD; Don Moore[5], PhD; Bonnie M Miller[1,5], MMHC, MD

[1]Center for Advanced Mobile Healthcare Learning, Vanderbilt University Medical Center, Nashville, TN, United States
[2]Department of Anesthesiology, School of Medicine, Vanderbilt University, Nashville, TN, United States
[3]Center for Artificial Intelligence, Research, and Evaluation, IBM Watson Health, Cambridge, MA, United States
[4]Clinical Evidence Development, Aetna Medical Affairs, CVS Health, Hartford, CT, United States
[5]Department of Medical Education and Administration, School of Medicine, Vanderbilt University, Nashville, TN, United States
[6]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN, United States
[7]Center of Excellence in Applied Artificial Intelligence, Vanderbilt University Medical Center, Nashville, TN, United States
[*]these authors contributed equally

Corresponding Author:
Kelly Jean Thomas Craig, PhD
Clinical Evidence Development
Aetna Medical Affairs
CVS Health
151 Farmington Avenue
RC31
Hartford, CT, 06156
United States
Phone: 1 970 261 3366
Email: craigk@aetna.com

Abstract

Background: The use of artificial intelligence (AI)–based tools in the care of individual patients and patient populations is rapidly expanding.

Objective: The aim of this paper is to systematically identify research on provider competencies needed for the use of AI in clinical settings.

Methods: A scoping review was conducted to identify articles published between January 1, 2009, and May 1, 2020, from MEDLINE, CINAHL, and the Cochrane Library databases, using search queries for terms related to health care professionals (eg, medical, nursing, and pharmacy) and their professional development in all phases of clinical education, AI-based tools in all settings of clinical practice, and professional education domains of competencies and performance. Limits were provided for English language, studies on humans with abstracts, and settings in the United States.

Results: The searches identified 3476 records, of which 4 met the inclusion criteria. These studies described the use of AI in clinical practice and measured at least one aspect of clinician competence. While many studies measured the performance of the AI-based tool, only 4 measured clinician performance in terms of the knowledge, skills, or attitudes needed to understand and effectively use the new tools being tested. These 4 articles primarily focused on the ability of AI to enhance patient care and clinical decision-making by improving information flow and display, specifically for physicians.

Conclusions: While many research studies were identified that investigate the potential effectiveness of using AI technologies in health care, very few address specific competencies that are needed by clinicians to use them effectively. This highlights a critical gap.

(JMIR Med Inform 2022;10(11):e37478) doi:10.2196/37478

KEYWORDS

artificial intelligence; competency; clinical education; patient; digital health; digital tool; clinical tool; health technology; health care; educational framework; decision-making; clinical decision; health information; physician

Introduction

Artificial intelligence (AI), defined as the "branch of computer science that attempts to understand and build intelligent entities, often instantiated as software programs," [1] has been applied in the health care setting for decades. Starting in the 1960s, a cadre of computer scientists and physicians developed an interest group around AI in Medicine (AIM) [2]. By the time funding sources became aligned with opportunities in the 1980s, AI was in its "expert system" era, using rules and knowledge derived from human experts to solve problems, primarily related to medical diagnosis [3]. Projects that developed these knowledge-based systems resulted in the creation of valuable information infrastructures, including standards, vocabularies, and taxonomies that continue to anchor electronic health records (EHR) [4]. Rule-based clinical decision support (eg, case-specific clinical alerts) is an important component of today's EHR, but it is no longer considered to be true AI [5].

Since these early forays into AI, great progress has been made in the structure and scope of information and computing technologies, as well as in data and computational resources, enabling the development of a much more powerful generation of AI tools. Human-machine collaborations exploiting these tools are already evident across professional health care practice. The ubiquitous use of personal computers and smartphones linked to external databases and highly connected AI-driven networks supports individual, team, and health system performance. This powerful new generation of AI-based tools will have wide-ranging impacts on the entire health care ecosystem, but concerns about potentially serious technical and ethical liabilities have also emerged [6].

Despite inevitable challenges, all those engaged in the practice and administration of health care should prepare for a future shaped by the presence of increasingly intelligent technologies, including robotic devices, clinical decision support systems based on machine-learning algorithms, and the flow of data and information from multiple sources, ranging from health information technology systems to individual patient sensors. While the health care and health professions education community are perched on the forefront of these complex developments, like many organizations, they may not be prepared to recognize and adequately respond to the deep-change indicators of next-generation technologies [7]. Eaneff and others recently called for new administrative infrastructures to help manage and audit the deluge of AI-induced change [8]. It is imperative for educators to be a part of that infrastructure—to actively engage in deliberations about intended changes in the working-learning environment—so that implications for learning and the needs of learners will be considered as a part of any change management process.

This impending onslaught also creates an urgent mandate for health care organizations, educators, and professional groups to consider the range of professional competencies needed for the effective, ethical, and compassionate use of AI in health care work. While numerous authors have called for structured and intentional learning programs, to date, there has been no published framework to guide teaching, learning, and assessing health care students and practitioners in this emerging and transformative domain [7,9-12]. Additionally, while there are many accredited programs (including board certification) in clinical informatics, they are focused on developing, implementing, and managing AI-based tools. However, these programs do not provide competencies for noninformatics users of AI-based tools, which represents a large gap in knowledge.

To inform these critical needs, this study aimed to systematically identify research studies that reported on provider competencies and performance measures related to the use of AI in clinical settings.

Methods

Study Design

A scoping review was conducted in accordance with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [13,14] with an *a priori* protocol. The objective was to systematically identify studies that specify competencies and measure performance related to the use of AI by health care professionals. Studies had to include students or postgraduate trainees in clinical education settings across medicine, nursing, pharmacy, and social work, or practicing clinicians participating in professional development activities.

Search Strategy

A systematic search query of MEDLINE via PubMed, CINAHL, and the Cochrane Library was conducted to identify references published or available online between January 1, 2009, and July 22, 2020 (Tables S1 to S3 in Multimedia Appendix 1). Queries including medical subject headings (MeSH) and keywords were designed around the following PICOST (population, intervention, control, outcomes, study design, and time frame) framework: (1) populations under consideration included all participants in any phase of clinical education including faculty and health care worker professional development (eg, clinical education participants in medical, nursing, or pharmacy; medical faculty and professional development; health care, clinical, or medical social workers); (2) interventions focused on AI-based tools (eg, AI terms, precision medicine, decision-making, speech recognition, documentation, computer simulation, software, patient participation or engagement, patient monitoring, health information exchange, EHR, and cloud computing) used in all settings; (3) no comparisons were required; (4) outcomes included the identification of clinical competencies and their respective measurements or domains; (5) study settings and limits included those with an abstract, conducted in humans, designed as primary studies or systematic reviews (with the same inclusion criteria), took place in US settings, and were published in English language; and (6) time—the introduction of the Health Information Technology for Economic and Clinical Health Act of 2009 was a distinguishing time point for this protocol [15,16]. AI-related tool use increased dramatically because of the organizational changes needed to accommodate meaningful use of health information technology in clinical care, justifying 2009 as a logical start point for this review.

Notably, during the protocol generation and scoping of the literature, it was determined that the MeSH term "informatics" lowered the precision (ie, irrelevant records returned) of our search strategy and greatly expanded the scope of literature to be reviewed. As such, exploded terms (eg, retrieving results under that selected subject heading and all of the more specific terms listed below in the tree) under the MeSH term "medical informatics," including "health information exchange," and fully exploded terms under "medical informatics applications" were applied. MeSH terms including "decision-making," "computer-assisted," "decision support techniques," "computer simulation," "clinical information systems," and "information systems," were among the relevant terms used. Similarly, due to imprecision, "information technology" MeSH term and "digital health" keyword were substituted with specific relevant examples for this study. Please see the search strategies provided in Tables S1 to S3 in Multimedia Appendix 1, which were created to support this scoping review protocol.

Screening Process

Screening of each title and abstract and each full text was performed by a single reviewer for relevance against the inclusion/exclusion criteria (Table S4 in Multimedia Appendix 1).

Studies with a population exclusively limited to other types of clinicians, including allied health professionals (eg, dental hygienists, diagnostic medical sonographers, dietitians, medical assistant, medical technologists, occupational therapists, physical therapists, radiographers, respiratory therapists, and speech language pathologists), dentists, and counselors were excluded.

Relevant AI-based tools could be used in all settings (eg, outpatient, inpatient, ambulatory care, critical care, and long-term care) of clinical practice, and there was a focus on subsets that incorporated either machine learning, natural language processing, deep learning, or neural networking. Exclusions were made for AI-based tools that did not meet inclusion criteria, such as studies using technology that did not incorporate relevant AI-based tools, when the methods provided regarding the tool did not explicitly define what type of AI methodology is incorporated, or if the AI is not machine learning, natural language processing, deep learning, or neural networking. Studies on robotics (eg, robotic surgery) were excluded unless AI was a noted part of the technology.

To identify studies that specified competencies and measured performance related to the use of AI by health care professionals, the inclusion criteria (Table S4 in Multimedia Appendix 1) were limited to the 6 professional education domains of competence (eg, patient care, medical knowledge or knowledge for practice, professionalism, interpersonal and communication skills, practice-based learning and improvement, and systems-based practice) or Entrustable Professional Activities and performance. Studies were excluded if they did not report on competency-based clinical education to provide either an evaluation of a program and its outcomes related to learner achievement; a framework for assessing competency including a performance level (ie, appraisal) for each competency; or information related to instructional design, skills validation, or attitudes related to competency mastery.

The results were tracked in DistillerSR [17]. Additionally, a validated AI-based prioritization tool embedded in DistillerSR was used to support the single screening of titles and abstracts to modify or stop the screening approach once a true recall at 95% was achieved [18]. Studies had to specify competencies and measure performance related to the use of AI by health care professionals.

Data Extraction

Data were abstracted into standardized forms (Table S5 in Multimedia Appendix 1) for synthesis and thematic analysis by 1 reviewer, and the content was examined for quality and completeness by a second reviewer, assuring that each included manuscript was dually reviewed. Abstraction for clinical education outcomes focused on how the necessary clinician competencies were described and measured. Conflict resolution was provided by consensus agreement.

Study Quality

Study quality was assessed by dual review using the Oxford levels of evidence [19].

Results

Search Outcomes

Literature searches yielded 3476 unique citations (Figure 1), of which 109 (3.14%) articles were eligible for full-text screening. Upon full-text screening, 4 articles met our inclusion criteria [20-23]. Abstractions of the included studies can be found in Tables 1 and 2 and Table S5 in Multimedia Appendix 1.

Figure 1. Results of literature search, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram [14]. Summary of articles identified by systematic search queries and tracking of articles that were included and excluded across the study screening phases with reasons for exclusion of full texts provided. AI: artificial intelligence.

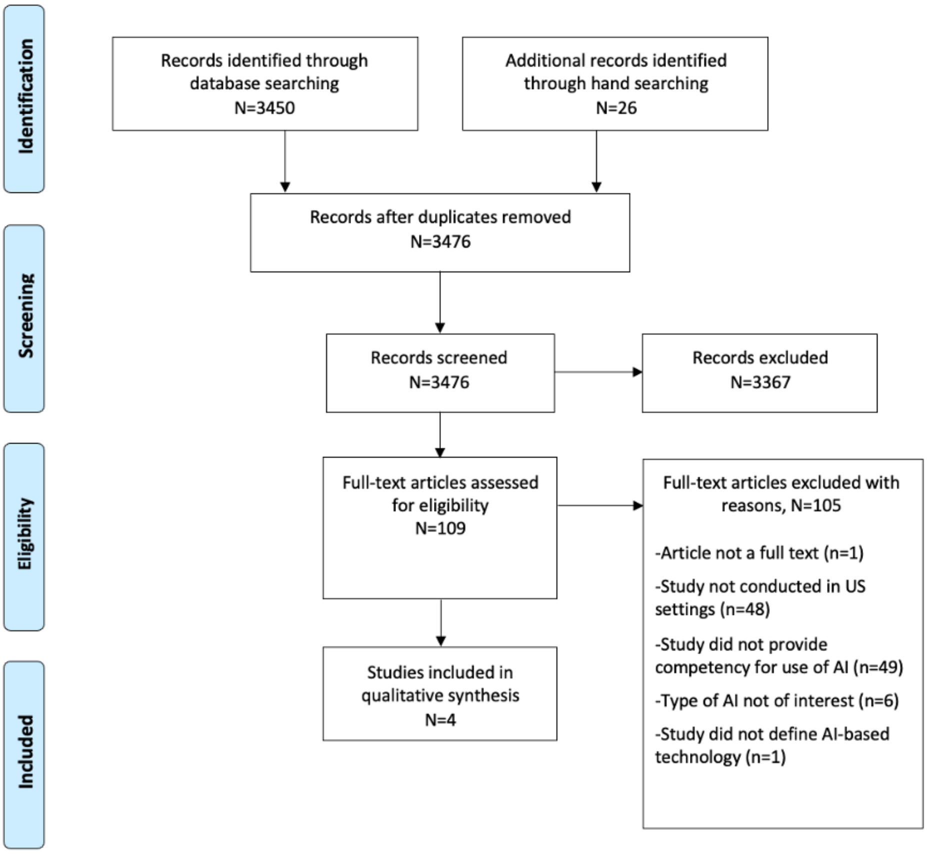

Identification
Screening
Eligibility
Included
Records identified through database searching N=3450
Additional records identified through hand searching N=26
Records after duplicates removed N=3476
Records screened N=3476
Records excluded N=3367
Full-text articles assessed for eligibility N=109
Full-text articles excluded with reasons, N=105
-Article not a full text (n=1)
-Study not conducted in US settings (n=48)
-Study did not provide competency for use of AI (n=49)
-Type of AI not of interest (n=6)
-Study did not define AI-based technology (n=1)
Studies included in qualitative synthesis N=4

Table 1. Summary of study characteristics: design and population.

Ref. No.	Ref., Year	Design; level of evidence[a]	Clinical setting	Users of AI[b]	Stage of clinical education	Stage of clinical use	Total, n (% male)	Age (years), race or ethnicity (%)	Study duration or follow-up
1	Bien, 2018 [23]	Modeling and evaluation; *2b*[c]	Large academic hospital; imaging department	Orthopedic surgeons; general radiologists	Practicing physicians	Implementation	N/R[d] (N/R)	N/R (N/R)	N/R
2	Hirsch, 2015 [22]	Evaluation; *4*[e]	Large private hospital; large academic medical center; nephrology and internal medicine departments	Internal medicine physicians; nephrologists	Graduate medical education (internal medicine residents and interns; nephrology fellows)	Implementation	12 (N/R)	N/R (N/R)	~9 months
3	Jordan, 2010 [21]	Evaluation; *4*	Large academic hospital; cardiothoracic intensive care department	Intensive care unit nurses	Practicing nurses	Implementation	N/R (N/R)	N/R (N/R)	N/R
4	Sayres, 2019 [20]	Experimental 3-arm observational study; *2b*	Large academic hospitals, large health systems, and specialist office; ophthalmology department	Ophthalmologists	Practicing physicians	Implementation	10 (N/R)	N/R (N/R)	N/R

[a] Adapted from Oxford Levels of Evidence [19].

[b] AI: artificial intelligence.

[d] *Level 2b*: individual cohort, modeling, or observational studies.

[c] N/R: not reported.

[e] *Level 4*: case series or poor-quality cohort studies.

Table 2. Summary of study characteristics: clinical competency and performance assessment.

Ref. No.	Ref., Year	Professional education domains of competence	Description (implied or explicit) of competency	User-AI[a] interface training and description	Performance assessment
1	Bien, 2018 [23]	- Patient care—clinical skills	Implied in methods; improve image interpretation	Training N/R[b]; interface not described	Metric N/P[c]; evaluate if AI assistance improves expert performance in reading MRI[d] images
2	Hirsch, 2015 [22]	- Patient care—clinical skills	Implied in methods; improve summarization of longitudinal patient record and information processing in preparation for new patients	Training N/R; authenticated user queries the database for a patient and is provided with a visual summary of content containing all visit, note, and problem information	Questionnaire; evaluate time and efficiency in information processing for patient care
3	Jordan, 2010 [21]	- Communication - Patient care—clinical skills - Systems-based practice	Implied in methods; improve handovers in peri-operative patient care by reducing communication and informational errors	Training N/R; patient summarization and visualization tool are used as an overlay to the existing electronic patient record	Questionnaire; evaluate if AI-based tool performs better than physicians to provide clinical information and patient status in ICU[e] handovers
4	Sayres, 2019 [20]	- Patient care—clinical skills	Implied in methods; improve reader sensitivity and increase specificity of fundal images	Readers were provided training and similar instructions for use; interface not described	Metric N/P; evaluate if AI assistance increases severity grades in model predictions by assessing sensitivity and specificity of reader

[a] AI: artificial intelligence.

[b] N/R: not reported.

[c] N/P: not provided.

[d] MRI: magnetic resonance imaging.

[e] ICU: intensive care unit.

Study Characteristics

Of the 4 studies, 3 (75%) studies were published in the past 5 years, and all 4 of the included studies were conducted in large, academic hospitals [20,22,23]. All AI-based tools in these identified studies were in a mature implementation phase and were being evaluated with either practicing physicians, residents, fellows, or nurses [20-23]. All 4 studies were undertaken to characterize the performance of internally developed niche AI software systems when used by health care professionals in specific practice settings (Table 1) [20-23].

All AI-based tools examined in these identified studies aimed to enhance an existing process, create new efficiencies, improve an outcome, and ultimately reduce cost of care [20-23]. Two of the AI-based tools were built on natural language processing frameworks [21,22] and 2 were based on deep learning processes [20,23]. One of the studies provided decision support in interpreting magnetic resonance imaging exams of the knee [23], 1 on enhancing clinician performance in detecting diabetic retinopathy [20], 1 on expediting EHR review prior to patient encounters [22], and 1 on enhancing the quality of patient handovers in the intensive care unit [21]. These systems were evaluated with measures of user satisfaction, usability, and performance outcomes. Studies used either observational or minimally controlled cohort designs, in which performance of the human-AI dyad was compared to expert performance or generalist performance alone. Three studies indicated moderate success with the AI interventions [20,21,23], and 1 had a neutral result (Table S2 in Multimedia Appendix 1) [22].

The impact of advanced data visualization, computerized image interpretation, and personalized just-in-time patient transitions are described in all 4 studies [20-23]. Competencies observed for use of these AI systems fell within the Accreditation Council for Graduate Medical Education patient care and communication competency domains [24]. However, the specific competencies clinicians required to use these innovations most effectively were not clearly described. Only 1 of the studies mentioned any form of training [20]; 3 did not describe any skill development processes for learners. None of the studies specified any need for understanding of basic AI forms, and none described the background information clinicians received about the development, training, and validation of the tools (Table 2).

Study Quality

Using Oxford Levels of Evidence [19] to examine study quality to measure the extent that methodological safeguards (ie, internal study validity) against bias were implemented, 2 studies provided Level 2b evidence as modeling summarizations [20,23], and 2 studies provided Level 4 evidence [21,22]. The overall quality identified is moderate to low, as half of the curated evidence was classified as Level 4.

Discussion

Principal Findings

The volume of studies initially identified for our review confirms predictions about the growth of AI in health care. However, of these nearly 3500 articles, only 4 met the inclusion criteria. This result begs a few questions. Were our requirements overly rigorous or are the research gaps truly that numerous? Moreover, does this result reinforce concerns about a lack of organizational preparedness?

Failure to address user competencies was the most common reason for study exclusion. Many of the excluded studies compared AI tool performance with that of practicing clinicians (*human versus machine*), while others used simulations to demonstrate the potential of AI innovations to improve clinical outcomes. Only 4 research studies were identified in our search [20-23] that addressed professional competencies observed by this new AI landscape; however, none of the identified studies described new AI-related clinical competencies that had to be developed. The limited evidence derived from this review points to a large gap in adequately designed studies that identify competencies for the use of AI-based tools.

While many skills will be specific for the AI intervention being employed, these "questions of competence" are broader than the technical skills needed for use of any one AI tool or type of intelligent support [25]. All health professionals will interact with these types of technologies during their daily practice and should "know what they need to know" before using a new system. System characteristics will profoundly impact patient and clinician satisfaction as well as clinical recommendations, treatment courses, and outcomes, so health system leaders must also *know what to know* before adopting new technologies across entire health care delivery enterprises. Health care professionals at all levels have the educational imperative to articulate, measure, and iterate competencies for thriving in this evolving interface of smart technology and clinical care.

The implementation of AI into clinical workflows without sufficient education and training processes to apply the technology safely, ethically, and effectively in practice could potentially negatively impact clinical and societal outcomes. Real-world deployment of AI has caused harms due to data bias (eg, algorithms trained using biased or poor-quality data) and societal bias (eg, algorithmic output reflects societal biases of human developer) [6,26]. These biases can inflate prediction performance, confuse data interpretation, and exacerbate existing social inequities (eg, racial, gender, and socioeconomic status). These ethical considerations bring additional responsibilities and oversight of both AI-based tool implementation and its associated data to the clinical care team. The scalability of AI-based tools can also increase the scale of associated risks [8,10]. These difficulties and potential risks should be identified and understood proactively, and skills for clinicians to approach them must be included in any comprehensive training program.

The scarcity of competencies identified by this scoping review reiterates the need to develop and recommended professional competencies for the use of AI-based tools [27,28]. Ideally, these competencies should promote the effective deployment of AI in shared decision-making models that sustain or even enhance compassion, humanity, and trust in clinicians and clinical care [29]. Additionally, user-centered design (eg, more specifically, human-centered design to develop human-centric AI algorithms) should also be considered in the development of educational frameworks to support AI-related competencies required for all clinicians to use these tools effectively in clinical settings. In follow-up to this report, the authors carried out structured interviews with thought leaders to develop such a competency framework, which subsequently can be tested and iteratively refined within both simulated and authentic workplace experiences [30].

Strengths and Limitations

This scoping review has several strengths. First, this is a novel and rigorous synthesis that adhered to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards. Second, its search strategy was comprehensive and inclusive, using keywords and MeSH terms for trainee populations, settings, interventions, and outcomes that would uncover all potential accounts of currently available evidence. Moreover, the availability of these comprehensive searches will support other studies examining AI and clinical education. Third, this study included the multiple types of health care professionals who might receive training and education for the use of AI in the clinical environment.

Our results should be interpreted in the context of a few limitations. The inclusion of US-only sites limits generalizability to other global settings and health system structures. It also may have eliminated additional salient investigations, although we imagine that the dearth of US studies predicts a similar deficit from other countries. Further, due to the heterogeneity of identified interventions, it would not have been possible to compare one training approach to another. A quality assessment tool was intentionally employed, as we only planned to measure the extent that methodological safeguards (ie, internal validity) against bias were implemented. Alternatively, a risk of bias assessment would have offered a bias judgement (ie, estimation of intervention effects) on such a quality assessment, and judgement of the evidence may have shifted with this approach [31]. The search cutoff date is another limitation, as other evidence may have been published since May 2020. Other limitations include single screening of titles and abstracts, English language restriction, and exclusion of studies reported in gray literature, including conference abstracts. In addition, we excluded articles that investigated the development of robotics-assisted competencies and those that measured the impact of computer vision tools in supporting technical learning in real and simulated settings. Finally, we restricted studies to those that evaluated the use of clinical AI and excluded those supporting other learning processes, although we recognize that tools such as AI-augmented learning management systems will also become a growing part of the health professions education landscape.

Conclusions

While many research studies were identified that investigate the potential effectiveness of using AI technologies in health care, very few address specific competencies that are needed by clinicians to use them effectively. This highlights a critical gap.

Acknowledgments

The authors wish to acknowledge the conceptual contributions of Gretchen P Jackson and Kyu Rhee. This study was supported by a grant from by IBM Watson Health.

Authors' Contributions

KJTC was responsible for methodology, project administration, and supervision. KJTC, RR, and KVG contributed to the validation of the study. KJTC and KVG were responsible for writing—original draft. All authors contributed to the paper's conceptualization, formal analysis, and writing—review and editing.

Conflicts of Interest

KJTC was employed by IBM Corporation. KVG, LLN, DM, and BMM are employed by Vanderbilt University Medical Center. RR is employed by Vanderbilt University School of Medicine.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 112 KB - medinform_v10i11e37478_app1.docx]

References

1. Yu K, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018 Oct;2(10):719-731. [doi: 10.1038/s41551-018-0305-z] [Medline: 31015651]
2. Patel VL, Shortliffe EH, Stefanelli M, Szolovits P, Berthold MR, Bellazzi R, et al. The coming of age of artificial intelligence in medicine. Artif Intell Med 2009 May;46(1):5-17 [FREE Full text] [doi: 10.1016/j.artmed.2008.07.017] [Medline: 18790621]
3. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1994 Jan 01;1(1):8-27 [FREE Full text] [doi: 10.1136/jamia.1994.95236141] [Medline: 7719792]
4. Hammond W, Cimino J. Standards in biomedical informatics. In: Biomedical Informatics Health Informatics. New York, NY: Springer; 2006:265-311.
5. Kulikowski CA. Beginnings of artificial intelligence in medicine (AIM): Computational artifice assisting scientific inquiry and clinical art - with reflections on present AIM challenges. Yearb Med Inform 2019 Aug;28(1):249-256 [FREE Full text] [doi: 10.1055/s-0039-1677895] [Medline: 31022744]
6. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
7. Wiljer D, Hakim Z. Developing an artificial intelligence-enabled health care practice: Rewiring health care professions for better care. J Med Imaging Radiat Sci 2019 Dec;50(4 Suppl 2):S8-S14. [doi: 10.1016/j.jmir.2019.09.010] [Medline: 31791914]
8. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. JAMA 2020 Oct 13;324(14):1397-1398. [doi: 10.1001/jama.2020.9371] [Medline: 32926087]
9. Hodges BD. Ones and zeros: Medical education and theory in the age of intelligent machines. Med Educ 2020 Aug;54(8):691-693. [doi: 10.1111/medu.14149] [Medline: 32160340]
10. Masters K. Artificial intelligence in medical education. Med Teach 2019 Sep;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]
11. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: Systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
12. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Academic Medicine 2018;93(8):1107-1109. [doi: 10.1097/acm.0000000000002044]
13. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 2009 Jun;26(2):91-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2009.00848.x] [Medline: 19490148]
14. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
15. Blumenthal D. Wiring the Health System — Origins and Provisions of a New Federal Program. N Engl J Med 2011 Dec 15;365(24):2323-2329. [doi: 10.1056/nejmsr1110507]
16. Health Information Technology for Economic and Clinical Health (HITECH) Act. Health Information Privacy. URL: https://tinyurl.com/76uvzx6a [accessed 2022-11-02]
17. DistillerSR. URL: https://www.evidencepartners.com/ [accessed 2022-11-02]

18. Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening - impact on reviewer-relevant outcomes. BMC Med Res Methodol 2020 Oct 15;20(1):256 [FREE Full text] [doi: 10.1186/s12874-020-01129-1] [Medline: 33059590]
19. Levels of evidence. The Centre for Evidence-based Medicine. 2009. URL: http://www.cebm.net/blog/2009/06/11/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/ [accessed 2022-11-02]
20. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 2019 Apr;126(4):552-564 [FREE Full text] [doi: 10.1016/j.ophtha.2018.11.016] [Medline: 30553900]
21. Jordan D, Rose SE. Multimedia abstract generation of intensive care data: the automation of clinical processes through AI methodologies. World J Surg 2010 Apr;34(4):637-645. [doi: 10.1007/s00268-009-0319-5] [Medline: 20012610]
22. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, et al. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc 2015 Mar;22(2):263-274 [FREE Full text] [doi: 10.1136/amiajnl-2014-002945] [Medline: 25352564]
23. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018 Nov;15(11):e1002699 [FREE Full text] [doi: 10.1371/journal.pmed.1002699] [Medline: 30481176]
24. Edgar L, McLean S, Hogan SO, Hamstra S, Holmboe ES. The milestones guidebook. ACGME. 2020. URL: https://www.acgme.org/globalassets/milestonesguidebook.pdf [accessed 2022-11-02]
25. Hodges B, Lingard L. The Question of Competence: Reconsidering Medical Education in the Twenty-First Century. Ithaca, NY, US: Cornell University Press; 2012.
26. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Artificial Intelligence in Healthcare. New York, US: Academic Press; 2020:295-336.
27. 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]
28. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: A report from the national academy of medicine. JAMA 2020 Feb 11;323(6):509-510. [doi: 10.1001/jama.2019.21579] [Medline: 31845963]
29. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bull World Health Organ 2020 Apr 01;98(4):245-250 [FREE Full text] [doi: 10.2471/BLT.19.237198] [Medline: 32284647]
30. Russell LL, Patel M, Garvey KM, Craig KJT, Jackson GP, Moore D, et al. Probably want to know a bit more about the magic: competencies for the use of artificial intelligence tools by healthcare workers in clinical settings. In: Health Professions Education Research Day. Nashville, TN, US: Vanderbilt University School of Medicine; Dec 03, 2021.
31. Furuya-Kanamori L, Xu C, Hasan SS, Doi SA. Quality versus risk-of-bias assessment in clinical research. J Clin Epidemiol 2021 Jan 13;129(2):172-175 [FREE Full text] [doi: 10.1016/j.jclinepi.2020.09.044] [Medline: 33422267]

Abbreviations

AI: artificial intelligence
EHR: electronic health records
MeSH: medical subject headings
PICOST: population, intervention, control, outcomes, study design, and time frame
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 22.02.22; peer-reviewed by S Baxter, K Masters; comments to author 15.03.22; revised version received 09.05.22; accepted 25.10.22; published 16.11.22.

Please cite as:
Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM
Considering Clinician Competencies for the Implementation of Artificial Intelligence–Based Tools in Health Care: Findings From a Scoping Review
JMIR Med Inform 2022;10(11):e37478
URL: https://medinform.jmir.org/2022/11/e37478
doi:10.2196/37478
PMID:36318697

Original Paper

The Real-World Experiences of Persons With Multiple Sclerosis During the First COVID-19 Lockdown: Application of Natural Language Processing

Deborah Chiavi[1*], MSc; Christina Haag[1,2*], PhD; Andrew Chan[3], MD; Christian Philipp Kamm[3,4], MD; Chloé Sieber[1,2], MSc; Mina Stanikić[1,2], MD; Stephanie Rodgers[1], PhD; Caroline Pot[5], MD; Jürg Kesselring[6], MD; Anke Salmen[3], MD; Irene Rapold[1], PhD; Pasquale Calabrese[7], MD; Zina-Mary Manjaly[8,9], MD; Claudio Gobbi[10,11], MD; Chiara Zecca[10,11], MD; Sebastian Walther[12], MD; Katharina Stegmayer[12], MD; Robert Hoepner[3], MD; Milo Puhan[1], PhD; Viktor von Wyl[1,2], PhD

[1]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[2]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

[3]Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

[4]Neurocenter, Lucerne Cantonal Hospital, Lucerne, Switzerland

[5]Service of Neurology, Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

[6]Department of Neurology and Neurorehabilitation, Rehabilitation Centre Kliniken Valens, Valens, Switzerland

[7]Division of Molecular and Cognitive Neuroscience, University of Basel, Basel, Switzerland

[8]Department of Neurology, Schulthess Klinik, Zurich, Switzerland

[9]Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland

[10]Multiple Sclerosis Center, Department of Neurology, Neurocenter of Southern Switzerland, Ente Ospedaliero Cantonale, Lugano, Switzerland

[11]Faculty of Biomedical Sciences, Università della Svizzera Italiana (USI), Lugano, Switzerland

[12]Translational Research Center, University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland

[*]these authors contributed equally

Corresponding Author:
Viktor von Wyl, PhD
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84
Zurich, 8001
Switzerland
Phone: 41 44 63 46380
Email: viktor.vonwyl@uzh.ch

Abstract

Background: The increasing availability of "real-world" data in the form of written text holds promise for deepening our understanding of societal and health-related challenges. Textual data constitute a rich source of information, allowing the capture of lived experiences through a broad range of different sources of information (eg, content and emotional tone). Interviews are the "gold standard" for gaining qualitative insights into individual experiences and perspectives. However, conducting interviews on a large scale is not always feasible, and standardized quantitative assessment suitable for large-scale application may miss important information. Surveys that include open-text assessments can combine the advantages of both methods and are well suited for the application of natural language processing (NLP) methods. While innovations in NLP have made large-scale text analysis more accessible, the analysis of real-world textual data is still complex and requires several consecutive steps.

Objective: We developed and subsequently examined the utility and scientific value of an NLP pipeline for extracting real-world experiences from textual data to provide guidance for applied researchers.

Methods: We applied the NLP pipeline to large-scale textual data collected by the Swiss Multiple Sclerosis (MS) registry. Such textual data constitute an ideal use case for the study of real-world text data. Specifically, we examined 639 text reports on the experienced impact of the first COVID-19 lockdown from the perspectives of persons with MS. The pipeline has been implemented in Python and complemented by analyses of the “Linguistic Inquiry and Word Count” software. It consists of the following 5 interconnected analysis steps: (1) text preprocessing; (2) sentiment analysis; (3) descriptive text analysis; (4) unsupervised learning–topic modeling; and (5) results interpretation and validation.

Results: A topic modeling analysis identified the following 4 distinct groups based on the topics participants were mainly concerned with: “contacts/communication;” “social environment;” “work;” and “errands/daily routines.” Notably, the sentiment analysis revealed that the “contacts/communication” group was characterized by a pronounced negative emotional tone underlying the text reports. This observed heterogeneity in emotional tonality underlying the reported experiences of the first COVID-19–related lockdown is likely to reflect differences in emotional burden, individual circumstances, and ways of coping with the pandemic, which is in line with previous research on this matter.

Conclusions: This study illustrates the timely and efficient applicability of an NLP pipeline and thereby serves as a precedent for applied researchers. Our study thereby contributes to both the dissemination of NLP techniques in applied health sciences and the identification of previously unknown experiences and burdens of persons with MS during the pandemic, which may be relevant for future treatment.

(JMIR Med Inform 2022;10(11):e37945) doi:10.2196/37945

KEYWORDS

natural language processing; multiple sclerosis; COVID-19; nervous system disease; nervous system disorder; textual data; health data; patient data; topic modeling; sentiment analysis; linguistic inquiry; medical informatics; clinical informatics

Introduction

Recent innovations in natural language processing (NLP) techniques and software have resulted in the emergence of numerous conveniently accessible and open-source analytical tools for the efficient evaluation of free-text data [1-4]. Textual data constitute a rich source of information, allowing the capture of unique perspectives, experiences, and individual needs through a broad range of different sources of information (eg, health-related content and emotional tone) [5,6]. While larger positive emotion vocabulary is linked to more mental well-being and better physical health, larger negative emotion vocabulary is associated with distress and decreased physical health [7].

In health research, the increasing availability of “real-world data” in the form of, for example, written text, constitutes a promising avenue to gain valid insights into themes that concern persons with chronic diseases in everyday life and thus are key to tailor individual support [8-10]. Many studies rely on interview techniques to gain such insights [11-16]. While conducting interviews represents the “gold standard” for gaining qualitative insights into individual experiences and perspectives, they may not always be feasible to assess individuals on a large scale. Scalable methods, which are very well suited for standardized quantitative assessments, may instead miss important information because they consist of predetermined items. Surveys that include open-ended text assessments can therefore be an appropriate way to qualitatively explore individual-level experiences and perspectives on a large scale in real-world environments.

Concurrently, practical guidelines for applied researchers concerning processing and evaluation procedures for textual information at a magnitude that is not feasible for manual analyses seem to be lacking. Given the novelty of the NLP method in the field of health research, we aim to share our work and experience in this manuscript to support applied researchers in implementing the NLP method in their own research. Therefore, the high-level aims of this study pertain to the investigation of the feasibility, usability, and scientific value of an NLP pipeline applied to the exploration of important life topics and themes in a large sample of persons with multiple sclerosis (MS) collected during a major health crisis. This study aims to provide practical guidance for applied researchers and leverages textual data from 639 well-documented persons with MS who described their live experiences during the first COVID-19 lockdown in Switzerland, as well as the availability of easy-to-use open-source tools for NLP.

At the content level, we addressed several specific research questions. We aimed to (1) identify cluster groups of persons with MS based on reported COVID-19–related topics; (2) determine the emotional tone underlying participants’ text entries; and (3) describe persons allocated to the same cluster group. For validation purposes, our analysis results were complemented by including independently collected information from the same database and a critical review by experts from the clinical or epidemiological research field.

Methods

Setting and Context

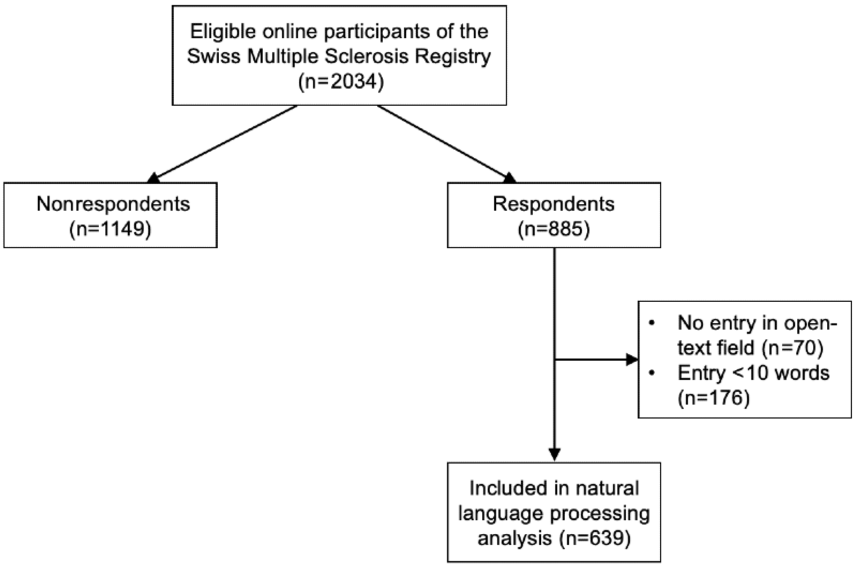
As laboratory-confirmed SARS-CoV-2 infections increased to up to almost 1500 cases daily (population size: 8.6 million inhabitants), the Swiss government implemented an initial lockdown between March 16 and April 27, 2020, to flatten the infection curve. On April 27, 2020, hairdressers, garden centers, flower shops, building supplies stores, and massage and beauty salons could reopen. In addition, entry requirements had been relaxed. On May 11, 2020, shops, restaurants, markets, libraries, and primary and secondary schools were reopened. The relaxations were accompanied by protection concepts. At the beginning of June 2020, all tourist facilities could open in compliance with protection measures. Events with up to 300 people could be held again, and gatherings with a maximum of 30 people were allowed again. On June 15, 2020, Switzerland lifted the entry regulations concerning all European Union/European Free Trade Association states and the United

Kingdom. On June 19, 2020, the Swiss Federal Council lifted the state of emergency. Most COVID-19 measures were lifted from June 22, 2020 (exception: large events with over 1000 people remained forbidden until the end of August 2020). All places open to the public needed to have a protection concept [17,18]. This first lockdown in Switzerland due to the COVID-19 pandemic resulted in pervasive and high levels of distress and isolation in the general population. These repercussions had a disproportionate effect on vulnerable subgroups of the population already burdened with pre-existing chronic diseases, such as MS. During the early stages of the pandemic, MS was also considered a risk factor for more severe COVID-19 symptoms, and persons with MS were advised to strictly adhere to preventive measures (ie, staying at home and keeping physical distance). At the end of April 2020, the lockdown measures were gradually lifted.

Data Sources

To assess the impact of the lockdown on the everyday lives of persons with MS, the Swiss MS Registry conducted a COVID-19–focused online survey among its over 2500 participants (Figure 1). The Swiss MS Registry is a nationwide survey-based registry encompassing adults with MS who reside in or receive MS-related care in Switzerland.

Figure 1. Flow diagram displaying the assessment procedure and subsequent selection procedure for online participants. Only online participants who described the experienced impact of COVID-19 on their personal life with at least 10 words were included in the text analysis.


Eligible online participants of the Swiss Multiple Sclerosis Registry (n=2034)
Nonrespondents (n=1149)
Respondents (n=885)
• No entry in open-text field (n=70)
• Entry <10 words (n=176)
Included in natural language processing analysis (n=639)

The “COVID-19 survey” was a brief online survey released by the Swiss MS Registry in response to the lockdown measures for the first wave, which assessed mental well-being and difficulties in accessing health care in times of COVID-19. The complete survey is provided in Multimedia Appendix 1. The COVID-19 survey starts with a short introduction, followed by a section on mental well-being, in which depressive symptoms are assessed using the Beck Depression Inventory FastScreen questionnaire [19]. This is followed by an assessment of physical well-being (ie, possible worsening of health or MS symptoms), fear of the presence of a serious illness (eg, coronavirus) in addition to MS, and perceived loneliness. The survey finally assesses general changes in individuals’ life situations due to the coronavirus. The open question, which is analyzed in the present, concerned the pandemic’s perceived impact on respondents’ daily lives. Specifically, participants were asked the following question: “How does the current coronavirus situation affect your personal life (eg, in terms of social contacts, everyday tasks, and health care provision)?” Participants were invited to document their answers without a maximum word limit in either German, French, or Italian (ie, the 3 official languages of the Swiss MS Registry). The COVID-19 survey was released online on April 10, 2020, and remained accessible until October 31, 2020. The current analysis includes all data collected until September 7, 2020.

For this study, the COVID-19 survey data were combined with sociodemographic and health-related data collected as part of the semiannual Swiss MS Registry assessments preceding the COVID-19 survey. Specifically, we employed the Self-Reported Disability Status Scale (SRDSS) to determine MS physical gait impairments. In this regard, the SRDSS classifies gait impairments based on 2 self-report questions that assess walking distance and the use of assistance devices [20]. Further, we determined health-related quality of life using the EuroQol 5-dimension scale (EQ-5D; index and visual analog scale) [21].

Ethics Approval

Approval has been obtained from the Cantonal Ethics Committee Zurich (PB-2016-00894). All participants enrolled in the Swiss MS Registry provided written (paper-pencil participants) or electronic (online participants) informed consent [22,23].

Descriptive Statistics

To characterize and compare online participants from the Swiss MS Registry participating in the COVID-19 survey with nonparticipants, sociodemographic and health characteristics were analyzed by means of N (%) for categorical data and medians (IQR) for continuous data. Descriptive statistics were based on the brief entry questionnaire, which is mandatory for all Swiss MS Registry participants and includes information on age, sex, MS type, diagnosis date, and any disease-modifying treatments.

Preprocessing and Analysis Pipeline for Free-Text Entries

This research implemented and evaluated a preprocessing and analysis pipeline to characterize and cluster free-text entries. To this end, we applied this pipeline to free-text entries about the impact of COVID-19 on the everyday lives of persons with MS. The entries were collected as part of the Covid-19 survey. The text preprocessing and analysis pipeline to be examined in this research consists of the following 5 interlinked consecutive steps: (1) text preprocessing; (2) descriptive text analysis; (3) sentiment analysis; (4) topic modeling; and (5) results interpretation and validation. An overview of the tools used in each step of the NLP pipeline can be found in Multimedia Appendix 2.

Step 1: Text Preprocessing

As the first step of the preprocessing procedure, Italian and French texts were translated into German using “DeepL Pro” [24], a tool for automatic text translation. Initially, we specified a cutoff for the minimum number of words for a text entry to be considered in the subsequent pipeline. As there are no generally valid guidelines applicable for our research in this regard, we based our decision on prior screening of the text entries and determined 10 words as cutoff to ensure sufficient informative content for the research question that we were interested in. Translation accuracy was checked manually and found to be very high. Further, punctuations and stop words (ie, common words without specific meaning like “the”) were removed using a publicly available German stop word list [25]. The remaining words were lemmatized (ie, changed to their root such as “studies” to “study”). Words not listed in dictionaries were converted into generic terms (eg, “Skype” to “video call”). This part of the pipeline was implemented using the Python library “spaCy” (version 2.3.2) [26].

Step 2: Descriptive Text Analyses

The second step of the pipeline concerned descriptive text analyses that involved determination of word frequencies as well as their visualization. For word frequency visualization, “word clouds” were compiled, which position all words into a graph where their relative size is determined by their overall frequency (ie, more frequent words are displayed larger in the plot) using the Python library “Wordcloud” (version 1.7.0) [27].

Step 3: Sentiment Analysis

The next step in pipeline pertained to the determination of linguistic indicators of overall text emotionality through sentiment analysis. To this end, 2 different text analysis resources were used: the well-established text analysis software “Linguistic Inquiry and Word Count” (LIWC) and further “SentimentWortschatz” (short “SentiWS”), a publicly available German-language resource for sentiment analysis. Sentiment analysis implemented in LIWC involved determining the text entries’ overall “emotional tone.” [28] “Emotional tone” is a summary variable provided by LIWC and represents the overall emotional coloration of a text. Scores range from 0 (negative tone) to 100 (positive tone), where a score of 50 indicates an even balance between positive and negative emotion words. Furthermore, we quantified text-based emotionality through “polarity scores” using the SentimentWortschatz sentiment-analysis resource (“SentiWS”) [29]. Polarity scores computed by SentiWS assess whether a word has a positive or negative connotation, ranging between −1 and 1. They are computed through a dictionary-based scoring algorithm that identifies words reflecting a negative or positive emotion. The SentiWS dictionary does not contain any polarity “shifters” or “intensifiers,” that is, words with an amplifying function, which weaken, intensify, or even reverse the meaning of an emotional word (eg, “not happy” or “very happy”). Since such amplifying words are key to accurately determine the polarity of a sentence, a German-language extension dictionary was used.

Step 4: Unsupervised Machine Learning–Topic Modeling

The final step of the pipeline concerns the implementation of “topic modeling,” which is an unsupervised text classification method with the aim to identify distinct clusters of common topics underlying free text (ie, underlying participants’ text entries) [30]. To determine distinct topic clusters, we implemented nonnegative matrix factorization, which is a topic modeling approach based on dimension reduction. Such dimension-reduction models are based on understanding a text corpus as a compilation of term frequencies. Nonnegative matrix factorization is based on a “bag of words” model, where text elements are represented in an unordered fashion. We further worked with unigrams, which means that each word corresponds to a text element (contrary to, for example, a bigram where a text element consists of 2 consecutive words). The reason for this methodological decision is that the majority of the words in the present data are meaningful in themselves in terms of co-occurrence and frequency.

We implemented this step using the Python libraries “scikit-learn” and “gensim” [31,32]. To determine the most suitable solution in terms of the number of distinct topics, we used the commonly used coherence score “C_v” as a criterion. “C_v” ranges from 0 (no topic coherence) to 1 (complete topic coherence). “C_v” scores for a modeling solution with 1 to 30 distinct topics are presented in Multimedia Appendix 3. We also computed the coherence score “UMass” but based the final topic modeling solution on “C_v” as it has been shown to be more appropriate for text data consisting of few words [33]. For sensitivity purposes, we repeated our analysis based on all available entries (ie, without word count restriction) in order to verify that topic clusters were stable.

Step 5: Results Interpretation and Validation

Finally, we labeled each of the distinct topic clusters with the term that occurred most often within the specific topic cluster. To further characterize individuals allocated to the distinct topic clusters, we compared independently collected sociodemographic measures across the groups through descriptive analyses. Given the descriptive nature of this research, we present 95% CIs instead of *P* values. We further linked emotional tone to the SRDSS score and years since diagnosis, which were both assessed as part of the previous biannual registry surveys. We also calculated the associations between emotional tone and the occurrence of new symptoms, the worsening of old symptoms, the presence of depressive symptoms, and the feeling of loneliness. For associations between interval-scaled variables, we calculated the Pearson correlation coefficient. For associations with ordinal variables, we computed the Spearman correlation coefficient. For correlations between interval-scaled and binary variables, we calculated the biserial point correlation coefficient. All associations were computed using the R package "psych" [34]. CIs for the Spearman correlation coefficient were computed using the R package "DescTools" [35]. Finally, the findings were critically reviewed by a team of experts coauthoring this study. The experts' backgrounds and specialist knowledge include neurology, neuropsychology, and epidemiology, as well as a personal health history of MS.

Results

Sample Characteristics

A total of 885 Swiss MS Registry participants (44.5% of all participants) completed a questionnaire pertaining to COVID-19 (Figure 1). As presented in Table 1, COVID-19–related survey respondents had a median age of 48 years, 70.3% (622/885) were female, and 67.9% (601/885) had relapsing-remitting MS (that is, with intermittent recovery of acute MS symptoms as opposed to continuously worsening primary and secondary progressing MS). Overall, participants who completed the COVID-19 survey were similar to nonparticipants (n=1149) in terms of their baseline characteristics (median age 47 years, 72.6% [834/1149] female, and 66.9% [769/1149] relapsing-remitting MS). From the overall sample of available survey responses (n=885; study flow chart provided in Figure 1), this study focused on entries of at least 10 words (n=639; Figure 2A). As there are no generally valid guidelines applicable for our research in this regard, we based our decision on prior screening of the text entries and determined 10 words as cutoff to ensure sufficient informative content for the research question that we were interested in. From this data source, 639 entries were used for the text analyses in this study.

The following sections describe the results obtained from the text preprocessing and analysis pipeline, which was applied to a sample of 639 COVID-19–related text entries provided by the Swiss MS Registry participants. The rationale for the methodological decisions of this study is provided in the Methods section.

Table 1. Description of Swiss Multiple Sclerosis Registry online participants and nonparticipants.

Characteristic[a]	Nonparticipants (did not complete the COVID-19 survey; N=1149)	Participants (completed the COVID-19 survey; N=885)
Age		
Value (years), median (IQR)	47.0 (38-56)	48.0 (39-56)
Missing information, n (%)	50 (4.4)	25 (2.8)
Gender, n (%)		
Female	834 (72.6)	622 (70.3)
Male	315 (27.4)	262 (29.6)
Missing information	0 (0)	1 (0.1)
Language, n (%)		
German	903 (78.6)	695 (78.5)
French	206 (17.9)	153 (17.3)
Italian	40 (3.5)	37 (4.2)
MS[b] type, n (%)		
CIS[c]	31 (2.7)	16 (1.8)
PPMS[d]	99 (8.6)	94 (10.6)
RRMS[e]	769 (66.9)	601 (67.9)
SPMS[f]	134 (11.7)	142 (16.0)
Transition between 2 MS types or unspecified	30 (2.6)	27 (3.1)
Missing information	86 (7.5)	5 (0.6)
Disease-modifying MS medication (immunotherapy), n (%)		
Yes	285 (24.8)	586 (66.2)
No	188 (16.4)	222 (25.1)
Missing information	676 (58.8)	77 (8.7)
Disease duration		
Value (years), median (IQR)	10.0 (5-18)	10.0 (4-17)
Missing information, n (%)	104 (9.1)	34 (3.8)
VAS[g] (health-related QLS[h])		
Value, median (IQR)	77 (54-90)	80 (60-90)
Missing information	185 (16.1)	121 (13.7)
EQ-5D[i]		
Value, median (IQR)	68.3 (49-88)	69.1 (51-91)
Missing information	185 (16.1)	121 (13.7)

[a] Percentages were rounded and may thus not add up to 100%.

[b] MS: multiple sclerosis.

[c] CIS: clinically isolated syndrome.

[d] PPMS: primary progressive MS.

[e] RRMS: relapsing-remitting MS.

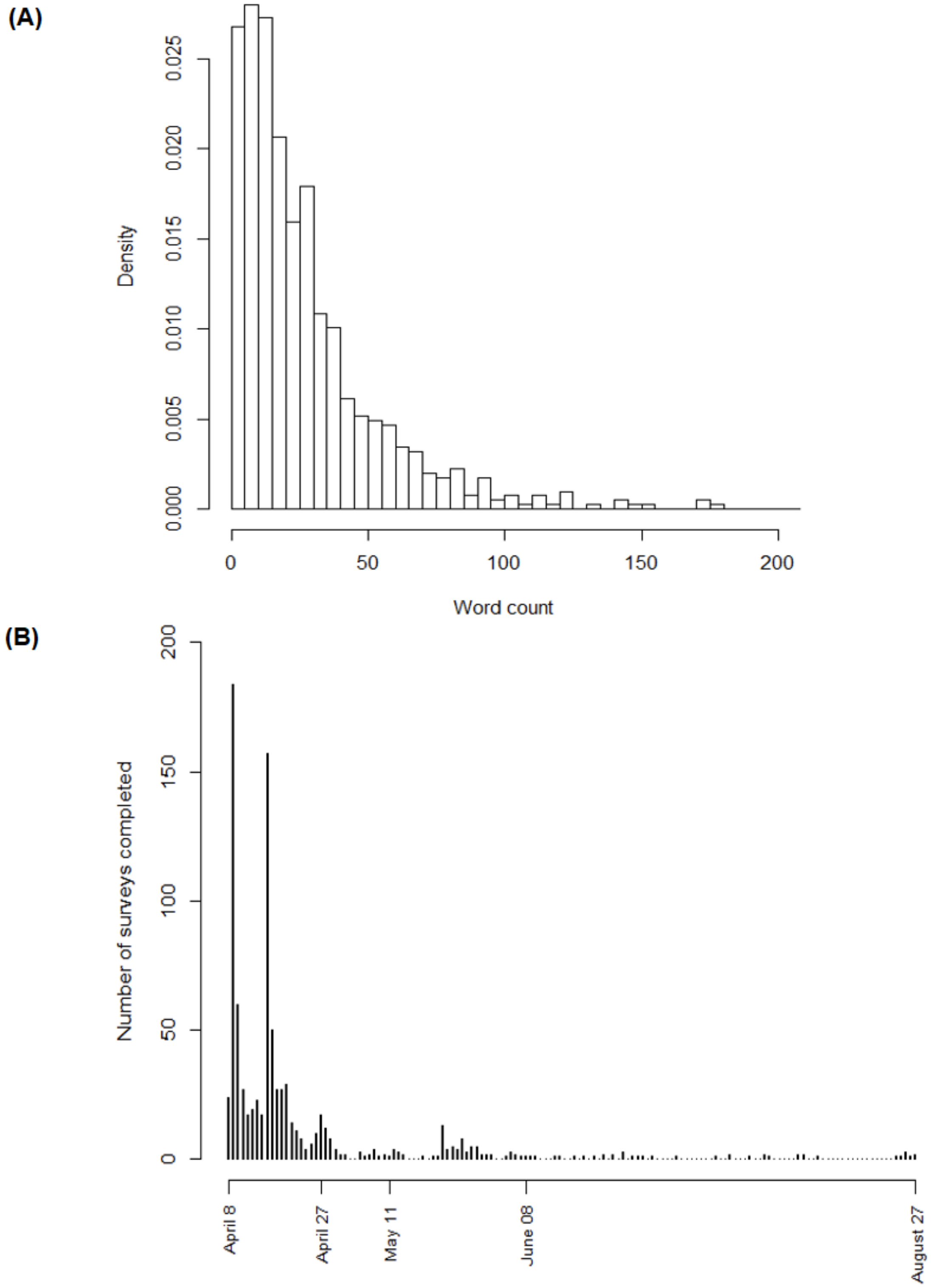
[f] SPMS: secondary progressive MS.

[g] VAS: visual analog scale.

[h] QLS: quality of life scale.

[i] EQ-5D: EuroQol 5-dimension scale.

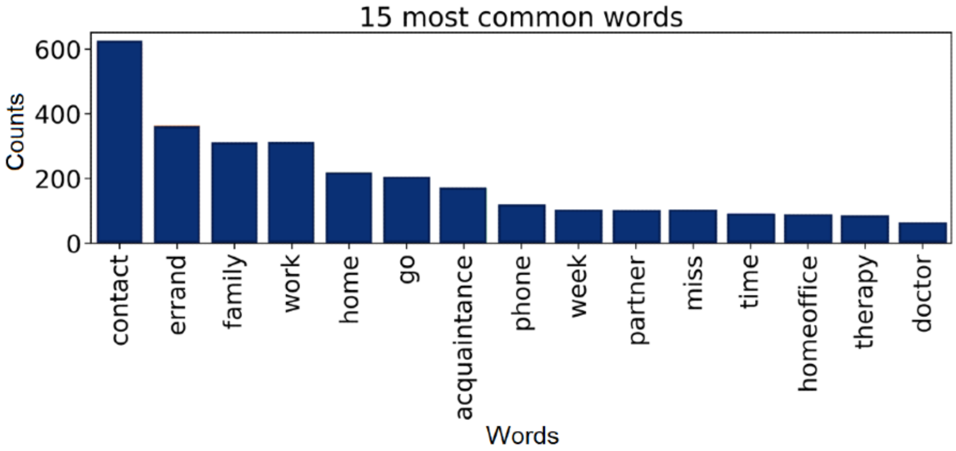
Figure 2. Survey responses included in this study. (A) Histogram depicting the text entries of different word lengths on the self-reported daily-life impact of COVID-19 (n=885). The number of words per text entry are plotted along the y-axis. (B) Amount of completed surveys across time (April 8, 2020, to August 27, 2020). Overall, 86.9% (555/639) of the responses were collected during the first lockdown (ie, before April 27, 2020). The number of completed surveys is displayed on the y-axis. Time (ie, days) is plotted along the x-axis.



Descriptive Text Analyses

Among all text responses used in this study, 86.9% (555/639) were collected during the first lockdown (before April 27, 2020; Figure 2B). In total, 80.1% (512/639) of these text entries were in German, 16.0% (102/639) in French, and 3.9% (25/639) in Italian. The median number of words per entry was 26 (IQR 16-44; following translation to German if necessary). Figure 3 visualizes the 15 most frequent keywords across the sample of text entries examined in this research. The most frequent words were “contact” (n=621), “errand” (n=364), “family” (n=307), “work” (n=307), and “home” (n=220).

Figure 3. Most frequent keywords across free-text descriptions on participants' perceived impact of COVID-19 on their personal life. Only text entries with at least 10 words in total were considered (n=639). "Stop words" (eg, "and" and "the") were removed prior to the analysis.



Sentiment Analysis

The possible full range of emotional tone of text entries ranged from 0 (negative) to 50 (neutral) up to 100 (positive). The mean emotional tone of participants' text entries was 34.7 (SD 37.7), thus reflecting an overall negative emotional tone. The distribution of emotional tone quartiles (1st quartile: 0-24; 2nd quartile: 25-49; 3rd quartile: 50-74; 4th quartile: 75-100) revealed that most of the 639 entries fell into the 1st quartile and thus were of overall negative quality (439/639, 68.7%). Importantly, most of the remaining text entries fell into the 4th quartile and thus were unambiguously of positive quality (160/639, 25.0%), while only few text entries were allocated to the intermediate quartiles (2nd quartile: 7/639, 1.1%; 3rd quartile: 33/639, 5.2%). The skewed distribution of the emotional tone of the participants' text entries explains the large standard deviation.

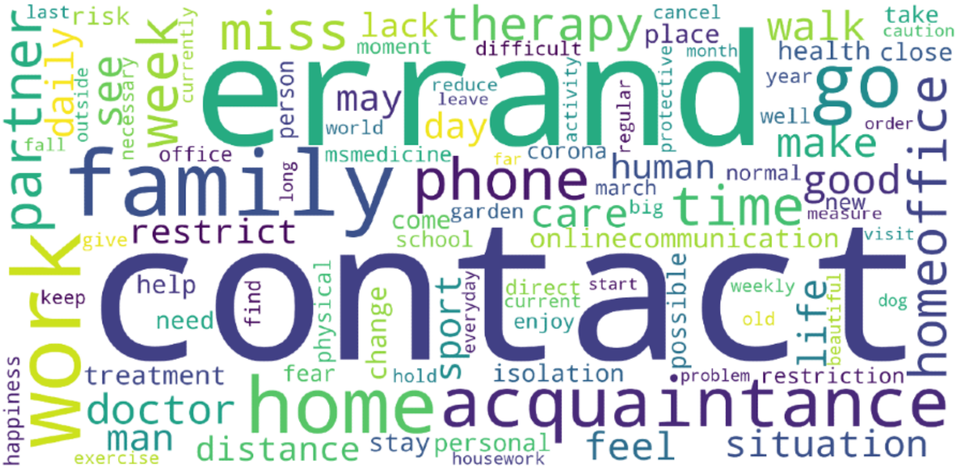
In terms of changes in COVID-19 measures across time, the average emotional tone across text entries did not differ during the lockdown (April 6 to 27; n=555; mean 35.32, SD 37.98; 95% CI 32.16-38.48) compared to the period during which restrictive measures were gradually lifted (April 28 to September 07; n=84; mean 30.58, SD 35.68; 95% CI 22.95-38.21).

Text-based polarity scores (ranging from −1 to 1) were comparable to those for emotional tone. Polarity scores were of overall negative valence (mean −0.10, SD 0.65), and 38.8% (248/639) of the entries had a polarity score below 0. Polarity scores based on text entries collected during first lockdown did not differ from those based on text entries collected during the time when measures were eased (following the lockdown; mean −0.13, SD 0.62).

Unsupervised Learning–Topic Modeling

Finally, the 639 text entries were grouped into distinct clusters through an unsupervised topic modeling procedure. Results revealed that a 4-group solution would be most suitable for the data structure. A word cloud visualizing the most frequent keywords related to the impact of COVID-19 on participants' personal lives across the complete study sample can be found in Figure 4. Word clouds for the 4 distinct topic groups are provided in Multimedia Appendix 4. The 4 distinct "topic groups" were labeled with the most frequent keywords (group 1: "contacts/communication," group 2: "social environment," group 3: "work," and group 4: "errands/daily routines"). A table characterizing the 4 distinct "topic groups" is provided in Multimedia Appendix 5. Text entries that were allocated to the "contacts/communication" group (group 1; 14.6% [119/639] of all text entries) captured how persons with MS experienced the contact restrictions. One of the most frequent words in this topic group was "miss." Importantly, text entries allocated to this group were of increasingly negative polarity. On the other hand, polarity scores in the "social environment" group (group 2; 21.4% [174/639] of all entries) and "work" group (group 3; 17.9% [146/639] of all entries) were more balanced. Finally, the "errands/daily routines" group (group 4; 24.5% [200/639] of all entries) included keywords that reflected daily routines (eg, "errands" and "going for a walk"). This group included the largest percentage of positive polarity scores (56.5%, 113/200). Repetition of the topic modeling analyses using all available text entries consistently found modeling 4 topic clusters to be ideal.

Figure 4. Word cloud visualizing the most frequent keywords related to the impact of COVID-19 on participants' personal lives across the complete study sample. Word size reflects the relative frequency of a specific word in comparison to the total number of analyzed words. Only text entries with at least 10 words in total were considered (n=639).



Sociodemographic and Health Characteristic Profiles

Additionally, we examined whether different sociodemographic and health characteristics were linked to distinct topic groups. The "contacts and communication" topic group tended to be older (median age: 49.5 years), live alone (27.7%, 33/119), be employed (second most; 63.9%, 76/119), and have lower levels of ambulatory disability (ie, persons who can move around without walking aids as measured with the self-reported disability scale [SRDSS], scores ranging between 0 and 3.5; 76.5%, 91/119). This group also reported the second highest health-related quality of life (median visual analog scale score: 80). Individuals allocated to the "social environment" topic group were more likely to have children (highest percentage; 50.6%, 88/174) who were typically under 18 years old (27.0%, 47/174). Further, pronounced mobility restrictions (ie, SRDSS scores greater than 3.5, thus requiring walking aids such as crutches or a wheelchair) were more frequent in this group, while health-related quality of life was comparatively lower (median EQ-5D: 0.65; median visual analog scale score: 75). Individuals allocated to the "work" topic group were most often employed compared to individuals in the other 3 topic groups (87.0%, 127/146), had SRDSS scores in the 0-3.5 range (highest proportion; 82.2%, 120/146), and had overall good quality of life (median EQ-5D index: 0.75; median visual analog scale score: 81). The "errands/daily routines" topic group had the most number of female research volunteers (79.0%, 158/200) and the highest proportion of persons on disability benefits (36.5%, 73/200). Quality of life in this group was higher as indicated by the visual analog scale (median score: 81). Finally, we examined the characteristics of online participants whose text entries had to be excluded as they were too short (n=176 entries). Individuals whose text entries had to be excluded were comparable to those of topic group 2 in terms of their sociodemographic characteristics (data not shown). Notably, the 3 most frequent keywords in the excluded entries (ie, "contacts," n=64; "errands," n=13; and "work," n=10) were also present in the 4 topic groups.

We further examined whether emotional tone was linked to measures of physical or mental well-being. Emotional tone was not linked to the SRDSS score (rho=−0.02, 95% CI −0.09 to 0.06; S=39575496, *P*=.69) or the number of years since the initial MS diagnosis (*r*=−0.03, 95% CI −0.11 to 0.05; t_{628}=−0.68333; *P*=.49). It was also not linked to the occurrence of new symptoms (*r*=−0.04, 95% CI −0.12 to 0.03; t_{633}=−1.121; *P*=.26) or the worsening of new symptoms (*r*=−0.07, 95% CI −0.14 to 0.01; t_{636}=−1.67; *P*=.09). However, emotional tone was significantly correlated with the presence of depressive symptoms (*r*=−0.10, 95% CI −0.19 to −0.02; t_{627}=−2.49; *P*=.01) and feelings of loneliness (*r*=−0.12, 95% CI −0.18 to −0.02; t_{630}=−2.92; *P*=.004). For all measures, less than 4% of the values were missing.

Discussion

Principal Findings

Here, we illustrate the application and subsequent evaluation of an NLP pipeline for the analysis of free-text data. Specifically, we applied this pipeline to text data on the experienced impact of the first COVID-19 lockdown from the perspectives of persons with MS collected by the Swiss MS Registry. Our study thus sheds light on individual daily-life experiences of the first COVID-19 lockdown in a vulnerable population.

In this study, we demonstrated both the feasibility and scientific value of an automated text preprocessing and NLP analysis pipeline based on existing open-source software in Python suitable for large-scale text data. The pipeline allows to preprocess real-world text data in an efficient fashion and to conduct timely and innovative analyses, including unsupervised machine learning. In light of a dearth of practical guidance for such real-world text data preprocessing and analysis procedures suitable for applied researchers, this pipeline has the potential to contribute to the dissemination of methodological knowledge, allowing to tap the potential of free-text data to capture individual perspectives and needs in health research. This study is embedded into the Swiss MS Registry, which is a large-scale well-documented longitudinal study. The registry's data thus constitute an optimal use case for the application and evaluation of such a pipeline and the broad range of available data sources allowed that characterize individuals allocated to the distinct topic cluster groups in terms of specific characteristics. This study demonstrates the potential of open-ended questions in complementing traditional standardized assessment methods to capture unexplored information from individuals' own words and thereby may spark new hypotheses and future avenues in health research. This type of language processing would essentially constitute a synergy between structured data collection and other forms of qualitative assessments, which tend to be more time-consuming in terms of processing and analysis (eg, interviews). Real-world data are afflicted with a broad range of challenges (eg, typos and dialect), which need elaborate consideration through text preprocessing to ensure the validity of subsequent complex analyses. Our study is thus timely and innovative in nature given its focus on key challenges when leveraging text data sources originating from a real-world setting through an efficient pipeline programmed in Python.

In terms of individual experiences of the first COVID-19 lockdown, the themes that concerned persons with MS most during the first COVID-19 lockdown differed substantially across study participants. Specifically, our study identified the following 4 distinct COVID-19–related topic groups, which participants could be assigned to based on their experiences: "contacts/communication" (group 1); "social environment" (group 2); "work" (group 3); and "errands/daily routines" (group 4). It is important to mention that between-group comparisons of sociodemographic and health-related characteristics corroborate the disparity of the 4 topic groups. This new topic-based approach to characterize persons with MS provides a novel perspective on individual experiences of the first COVID-19 lockdown and further highlights heterogeneity in terms of individual needs. To the best of our knowledge, there are no comparable in-depth studies researching the individually perceived impact of COVID-19 using participants own words. With regard to the overall emotional tone underlying the text entries, our findings revealed that most text entries reflected negative emotional states. This adds to research emphasizing the high burden of COVID-19–related restrictions for persons with MS given their prior vulnerability [12]. Further, from a methodological perspective, the context of our study was ideal for the identification of distinct topic commonalities of wide-ranging relevance as the spectrum of topics that participants were concerned with was confined. On the contrary, studies researching mundane everyday life situations of persons with MS are likely to identify considerably more diverse topics (with smaller population sizes per topic group), which results in the necessity of more data and participants, as suggested by an ongoing analysis of health diary entries collected before the COVID-19 pandemic from the same study population (manuscript in preparation).

In parallel with this finding, the 4 topic groups also differed in terms of the emotional tone underlying their text descriptions. It is important to mention that the emotional tone was determined through an independent analysis approach (sentiment analysis). A correlation analysis revealed that emotional tone was not associated with MS traits or measures of physical well-being, but with psychological well-being in the form of depressive symptoms and feelings of loneliness. This result suggests that "emotional tone" in this study primarily reflects emotions that are directly related to the content of the text and the individual's situation. The most negative entries occurred in topic groups whose text entries predominately pertained to contacts and communication themes (group 1). In the topic groups concerning social environment (group 2) and work (group 3), the underlying emotional tone was more balanced, while in the topic group pertaining to errands and daily routines (group 4), the entries' emotional tone was predominantly positive. This observed heterogeneity in emotional tonality underlying the reported experiences of the first COVID-19–related lockdown is likely to reflect differences in emotional burden, individual circumstances, and ways of coping with the pandemic, which is in line with previous research in this matter. For instance, a US telephone survey on persons with MS conducted during the first lockdown found that a higher perceived impact of the pandemic on individuals' self-reported psychological well-being was linked to a higher impact of MS symptoms on individuals' daily lives. Further, by conducting interviews, a recent study found that persons reporting no or even a positive impact of the pandemic on their lives tended to cope with the pandemic situation with active problem-focused strategies [11-13]. In terms of personal values, however, another study examining young persons with MS also reported perceived positive effects of the pandemic situation in the form of personal, relational, and existential growth [36]. Accordingly, participants allocated to the "contacts and communication" topic group made the highest number of negative text entries and reported the lowest quality of life (median). Taken together, these findings are foreground to the burdensome effects of the pandemic in terms of isolation, and reduction or even loss of social contact/activities and personal exchange in vulnerable individuals such as persons with MS. Based on the sociodemographic and disease characteristics of topic group 1, feelings of isolation appeared exacerbated in persons with MS who were comparatively less impaired or living alone. This finding might be related to the fact that persons with high disease burden are more accustomed to daily life restrictions compared to those with less impairments.

Limitations

Despite its notable strengths, the present research has some limitations, which merit consideration. First, there is a dearth of well-established guidelines for NLP that consider the specificities of health research. Consequently, the implementation of different text classification modeling approaches might have resulted in slightly divergent clusters and overarching topics. As such, to examine the robustness of our findings, we reanalyzed our data using the well-established Latent Dirichlet Allocation approach, which yielded similar patterns compared to those reported (not shown in this article) and thus corroborates the robustness of the presented results. Topic modeling further groups frequently co-occurring words into clusters (ie, "topics"). This method is suitable for identifying topics underlying large-scale text data in a data-driven fashion to thereby generate novel insights that might have been missed by standardized quantitative assessments. Our study does, however, not provide information to specifically tailor MS treatment to the needs of an individual person. The emotional tone indicates a general trend of the overall valence of a topic, while there may be variations at the individual level. Our findings have revealed experiences and burdens of persons with MS during the COVID-19 pandemic that may be relevant to future treatments or may provide insights for future research. Further limitations pertain to the generalizability of the findings of the sample population to the total population of persons with MS in Switzerland. Participants of this study constitute a subsample of the Swiss MS Registry's participants. The registry itself covers the diversity of the Swiss population of persons with MS in terms of a broad range of characteristics [37]. The participants of the MS Registry subsample who completed the "COVID-19 survey" were comparatively younger, less disabled, and residing more often in the German-speaking region of Switzerland than the nonparticipants of the registry. However, we did not find any indications for systematic differences between the linguistic regions. The translation of non-German text entries into German through an automated translation software is afflicted with the risk of potential mistranslations, misinterpretations, and biases. However, it is important to mention that both exploratory count comparison of the most frequent keywords and manual spot-checking were not suggestive of any systematic differences across languages.

Conclusion

We demonstrated the potential of a preprocessing and NLP analysis pipeline for large-scale text data and applied it to COVID-19–related data collected by the Swiss MS Registry, which constitutes an optimal use case for the pipeline. Above and beyond providing practical guidance for applied researchers, our study has implications for efficiently leveraging large-scale textual data in health care settings. Electronic health records and clinical notes have received increasing attention as rich sources of information, which are accessible through the application of NLP techniques [38-40].

Our study further demonstrates an approach that complements structured and standardized assessments through individual participant perspectives and hence provides ecologically valid information. We provide practical guidance for applied health researchers who wish to follow a similar approach by (1) demonstrating the processing and analysis process using large-scale real-world data and (2) providing a detailed description of the pipeline, which is based (apart from LIWC) on freely available open-source software. Interested researchers can follow both the entire process and the software we use. Given the novelty of the emerging NLP field, we are, in this way, contributing to the establishment of good practice standards and the dissemination of knowledge around NLP methodology among applied researchers, especially those from the health sciences.

Acknowledgments

We thank Oliver Widler who developed an initial version of the natural language processing pipeline. We further thank the members of the Swiss Multiple Sclerosis (MS) Registry "Begleitgruppe" who contributed to text analysis validation. We are also most grateful to all participants of the Swiss MS Registry who dedicated their time and thereby made a crucial contribution to this research. Further, our gratitude goes to the Swiss MS Society for funding the Swiss MS Registry and for the continuous support. We are also very thankful to those study participants who, above and beyond providing their data, contributed essentially to all aspects of the Swiss Multiple Sclerosis Registry, including study design and implementation. Finally, we would like to thank the members of the Swiss MS Registry data center at the Institute for Epidemiology, Biostatistics and Prevention, University of Zurich. The members of the Swiss MS Registry are as follows: Bernd Anderseck, Pasquale Calabrese, Andrew Chan, Claudio Gobbi, Roger Häussler, Christian P Kamm, Jürg Kesselring (President), Jens Kuhle (Chair of the Clinical and Laboratory Research Committee), Roland Kurmann, Christoph Lotter, Marc Lutz, Kurt Luyckx, Patricia Monin, Stefanie Müller, Krassen Nedeltchev, Caroline Pot, Milo A Puhan, Irene Rapold, Anke Salmen, Klaas Enno Stephan, Zina-Mary Manjaly, Claude Vaney (Chair of the Patient and Population Research Committee), Viktor von Wyl (Chair of the IT and Data Committee), and Chiara Zecca. The Swiss MS Registry is supported by the scientific advisory board of the Swiss MS Society.

This research was funded by the Swiss Multiple Sclerosis Society and by seed grants from the Digital Society Initiative and the Participatory Science Academy of the University of Zurich.

Conflicts of Interest

CPK has received honoraria for lectures as well as research support from Biogen, Novartis, Almirall, Bayer Schweiz AG, Teva, Merck, Sanofi Genzyme, Roche, Eli Lilly, Celgene, and the Swiss Multiple Sclerosis (MS) Society (SMSG). AS has received speaker honoraria and/or travel compensation for activities with Almirall Hermal GmbH, Biogen, Merck, Novartis, Roche, and Sanofi Genzyme, and research support from the Swiss MS Society, none related to this work. The employer Department of Neurology, Regional Hospital Lugano [EOC], Lugano, Switzerland received financial support for CZ and CG's speaking and educational, research, or travel grants from Abbvie, Almirall, Biogen Idec, Celgene, Sanofi, Merck, Novartis, Teva Pharma, and Roche. AC has received speaker/board honoraria from Actelion (Janssen/J&J), Almirall, Bayer, Biogen, Celgene (BMS), Genzyme, Merck KGaA (Darmstadt, Germany), Novartis, Roche, and Teva, all for hospital research funds. He received research support from Biogen, Genzyme, UCB, the European Union, and the Swiss National Foundation. He serves as associate editor of the European Journal of Neurology, is on the editorial board for Clinical and Translational Neuroscience, and serves as topic editor for the Journal of International Medical Research. RH has received honoraria from Janssen, Lundbeck, Mepha, and Neurolite. SW has received honoraria from Janssen, Lundbeck, Mepha, Neurolite, and Sunovion. MS reports employment by Roche from February 2019 to February 2020. KS has received honoraria from from Janssen, Lundbeck and Mepha.

Multimedia Appendix 1
COVID-19 survey of the Swiss Multiple Sclerosis Registry.
[PDF File (Adobe PDF File), 38 KB - medinform_v10i11e37945_app1.pdf]

Multimedia Appendix 2
An overview of the tools used in each step of the natural language processing pipeline.
[PNG File , 180 KB - medinform_v10i11e37945_app2.png]

Multimedia Appendix 3
Graph showing topic coherence scores (blue dots) for topic models on the experience of the first COVID-19 lockdown in persons with multiple sclerosis, with 1 to 30 distinct topics. The number of modeled topics is plotted along the x-axis. Coherence scores are plotted along the y-axis. Topic coherence refers to the semantic similarity of words allocated to a distinct topic and constitutes a key goodness of fit measure for topic models. The full possible range of coherence scores is between 0 (no topic coherence) and 1 (complete topic coherence). A 4-topic model provides the optimal modeling solution for the data as indicated by the highest coherence score.
[PNG File , 966 KB - medinform_v10i11e37945_app3.png]

Multimedia Appendix 4
Word clouds visualizing the most frequent keywords related to the impact of COVID-19 on volunteers’ personal lives presented separately for each of the 4 topic cluster groups. Word size reflects the relative frequency of a specific word in comparison to the total number of analyzed words in a topic group. Only text entries with at least 10 words in total were considered.
[PNG File , 1188 KB - medinform_v10i11e37945_app4.png]

Multimedia Appendix 5
Characterization of study participants assigned to the 4 topic groups “contacts/communication,” “social environment,” “work,” and “errands/daily routines”.
[DOCX File , 34 KB - medinform_v10i11e37945_app5.docx]

References

1. Cammel SA, De Vos MS, van Soest D, Hettne KM, Boer F, Steyerberg EW, et al. How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. BMC Med Inform Decis Mak 2020 May 27;20(1):97 [FREE Full text] [doi: 10.1186/s12911-020-1104-5] [Medline: 32460734]
2. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
3. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
4. Mascio A, Kraljevic Z, Bean D, Dobson R, Stewart R, Bendayan R, et al. Comparative Analysis of Text Classification Approaches in Electronic Health Records. arXiv. 2005. URL: http://arxiv.org/abs/2005.06624 [accessed 2021-11-18]
5. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat. Lang. Eng 2017 Jan 30;23(5):649-685. [doi: 10.1017/S1351324916000383]
6. Elkin PL, Mullin S, Mardekian J, Crowner C, Sakilay S, Sinha S, et al. Using Artificial Intelligence With Natural Language Processing to Combine Electronic Health Record's Structured and Free Text Data to Identify Nonvalvular Atrial Fibrillation to Decrease Strokes and Death: Evaluation and Case-Control Study. J Med Internet Res 2021 Nov 09;23(11):e28946 [FREE Full text] [doi: 10.2196/28946] [Medline: 34751659]
7. Vine V, Boyd RL, Pennebaker JW. Natural emotion vocabularies as windows on distress and well-being. Nat Commun 2020 Sep 10;11(1):4525 [FREE Full text] [doi: 10.1038/s41467-020-18349-0] [Medline: 32913209]
8. Rivas R, Montazeri N, Le NX, Hristidis V. Automatic Classification of Online Doctor Reviews: Evaluation of Text Classifier Algorithms. J Med Internet Res 2018 Nov 12;20(11):e11141 [FREE Full text] [doi: 10.2196/11141] [Medline: 30425030]
9. Ferrario A, Demiray B, Yordanova K, Luo M, Martin M. Social Reminiscence in Older Adults' Everyday Conversations: Automated Detection Using Natural Language Processing and Machine Learning. J Med Internet Res 2020 Sep 15;22(9):e19133 [FREE Full text] [doi: 10.2196/19133] [Medline: 32866108]

10. Le Glaz A, Haralambous Y, Kim-Dufor D, Lenca P, Billot R, Ryan TC, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. J Med Internet Res 2021 May 04;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
11. Donisi V, Gajofatto A, Mazzi MA, Gobbin F, Busch IM, Ghellere A, et al. Insights for Fostering Resilience in Young Adults With Multiple Sclerosis in the Aftermath of the COVID-19 Emergency: An Italian Survey. Front Psychiatry 2020 Feb 22;11:588275 [FREE Full text] [doi: 10.3389/fpsyt.2020.588275] [Medline: 33692703]
12. Morris-Bankole H, Ho AK. The COVID-19 Pandemic Experience in Multiple Sclerosis: The Good, the Bad and the Neutral. Neurol Ther 2021 Jun 15;10(1):279-291 [FREE Full text] [doi: 10.1007/s40120-021-00241-8] [Medline: 33855692]
13. Talaat F, Ramadan I, Aly S, Hamdy E. Are multiple sclerosis patients and their caregivers more anxious and more committed to following the basic preventive measures during the COVID-19 pandemic? Mult Scler Relat Disord 2020 Nov;46:102580 [FREE Full text] [doi: 10.1016/j.msard.2020.102580] [Medline: 33296977]
14. Vogel AC, Schmidt H, Loud S, McBurney R, Mateen FJ. Impact of the COVID-19 pandemic on the health care of >1,000 People living with multiple sclerosis: A cross-sectional study. Mult Scler Relat Disord 2020 Nov;46:102512 [FREE Full text] [doi: 10.1016/j.msard.2020.102512] [Medline: 32977074]
15. Manacorda T, Bandiera P, Terzuoli F, Ponzio M, Brichetto G, Zaratin P, et al. Impact of the COVID-19 pandemic on persons with multiple sclerosis: Early findings from a survey on disruptions in care and self-reported outcomes. J Health Serv Res Policy 2021 Jul 18;26(3):189-197 [FREE Full text] [doi: 10.1177/1355819620975069] [Medline: 33337256]
16. Colais P, Cascini S, Balducci M, Agabiti N, Davoli M, Fusco D, et al. Impact of the COVID-19 pandemic on access to healthcare services amongst patients with multiple sclerosis in the Lazio region, Italy. Eur J Neurol 2021 Oct 14;28(10):3403-3410 [FREE Full text] [doi: 10.1111/ene.14879] [Medline: 33896086]
17. Easing and tightening of nationwide measures. Federal Office of Public Health. 2020 Nov 2. URL: https://www.bag.admin.ch/dam/bag/en/dokumente/mt/k-und-i/aktuelle-ausbrueche-pandemien/2019-nCoV/covid-19-tabelle-lockerung.pdf.download.pdf/Easing_of_measures_and_possible_next_steps.pdf [accessed 2022-10-19]
18. Coronavirus: Measures and ordinances. Federal Office of Public Health. 2022 May 4. URL: https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/massnahmen-des-bundes.html [accessed 2022-10-19]
19. Kliem S, Mößle T, Zenger M, Brähler E. Reliability and validity of the Beck Depression Inventory-Fast Screen for medical patients in the general German population. J Affect Disord 2014 Mar;156:236-239. [doi: 10.1016/j.jad.2013.11.024] [Medline: 24480380]
20. Kaufmann M, Salmen A, Barin L, Puhan MA, Calabrese P, Kamm CP, Swiss Multiple Sclerosis Registry (SMSR). Development and validation of the self-reported disability status scale (SRDSS) to estimate EDSS-categories. Mult Scler Relat Disord 2020 Jul;42:102148 [FREE Full text] [doi: 10.1016/j.msard.2020.102148] [Medline: 32371376]
21. Hinz A, Klaiberg A, Brähler E, König HH. [The Quality of Life Questionnaire EQ-5D: modelling and norm values for the general population]. Psychother Psychosom Med Psychol 2006 Feb 10;56(2):42-48. [doi: 10.1055/s-2005-867061] [Medline: 16453241]
22. Puhan MA, Steinemann N, Kamm CP, Müller S, Kuhle J, Kurmann R, Swiss Multiple Sclerosis Registry (SMSR). A digitally facilitated citizen-science driven approach accelerates participant recruitment and increases study population diversity. Swiss Med Wkly 2018 May 16;148:w14623 [FREE Full text] [doi: 10.4414/smw.2018.14623] [Medline: 29767828]
23. Steinemann N, Kuhle J, Calabrese P, Kesselring J, Disanto G, Merkler D, Swiss Multiple Sclerosis Registry (SMSR). The Swiss Multiple Sclerosis Registry (SMSR): study protocol of a participatory, nationwide registry to promote epidemiological and patient-centered MS research. BMC Neurol 2018 Aug 13;18(1):111 [FREE Full text] [doi: 10.1186/s12883-018-1118-0] [Medline: 30103695]
24. DeepL Translator. DeepL. URL: https://www.deepl.com/translator [accessed 2022-10-19]
25. Stopwords ISO. Gene Diaz. 2020. URL: https://github.com/stopwords-iso/stopwords-de [accessed 2022-10-19]
26. Industrial-strength natural language processing in Python. spaCy. 2022. URL: https://spacy.io/ [accessed 2022-10-19]
27. Oesper L, Merico D, Isserlin R, Bader GD. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. Source Code Biol Med 2011 Apr 07;6(1):7 [FREE Full text] [doi: 10.1186/1751-0473-6-7] [Medline: 21473782]
28. Meier T, Boyd RL, Pennebaker JW, Mehl MR, Martin M, Wolf M, et al. “LIWC auf Deutsch”: The Development, Psychometrics, and Introduction of DE- LIWC2015. PsyArXiv Preprints. URL: https://psyarxiv.com/uq8zt/ [accessed 2022-10-19]
29. Remus R, Quasthoff U, Heyer G. SentiWS - A Publicly Available German-language Resource for Sentiment Analysis. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010 Presented at: Seventh International Conference on Language Resources and Evaluation; May 2010; Valletta, Malta URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/490_Paper.pdf
30. Blei D, Carin L, Dunson D. Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis. IEEE Signal Process Mag 2010 Nov 01;27(6):55-65 [FREE Full text] [doi: 10.1109/MSP.2010.938079] [Medline: 25104898]
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011;12:2825-2830.

32. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010 May 22 Presented at: LREC 2010; 17-23 May, 2010; Valletta, Malta p. 45-50 URL: http://is.muni.cz/publication/884893/en
33. Röder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. In: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015 Presented at: Eighth ACM International Conference on Web Search and Data Mining; February 2-6, 2015; Shanghai, China p. 399-408. [doi: 10.1145/2684822.2685324]
34. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. R Project homepage. 2022. URL: https://CRAN.R-project.org/package=psych [accessed 2022-10-19]
35. Signorell A, Aho K, Alfons A, Anderegg N, Aragon T, Arachchige C, et al. DescTools: Tools for Descriptive Statistics. R Project. 2022 Sep 1. URL: https://cran.r-project.org/package=DescTools [accessed 2022-10-19]
36. Poli S, Rimondini M, Gajofatto A, Mazzi MA, Busch IM, Gobbin F, et al. "If You Can't Control the Wind, Adjust Your Sail": Tips for Post-Pandemic Benefit Finding from Young Adults Living with Multiple Sclerosis. A Qualitative Study. Int J Environ Res Public Health 2021 Apr 14;18(8):4156 [FREE Full text] [doi: 10.3390/ijerph18084156] [Medline: 33919974]
37. Kaufmann M, Puhan MA, Kuhle J, Yaldizli Ö, Magnusson T, Kamm CP, et al. A Framework for Estimating the Burden of Chronic Diseases: Design and Application in the Context of Multiple Sclerosis. Front Neurol 2019 Sep 4;10:953 [FREE Full text] [doi: 10.3389/fneur.2019.00953] [Medline: 31555205]
38. 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]
39. Yang X, Zhang H, He X, Bian J, Wu Y. Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models. JMIR Med Inform 2020 Dec 15;8(12):e22982 [FREE Full text] [doi: 10.2196/22982] [Medline: 33320104]
40. 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]

Abbreviations

EQ-5D: EuroQol 5-dimension scale
LIWC: Linguistic Inquiry and Word Count
MS: multiple sclerosis
NLP: natural language processing
SentiWS: SentimentWortschatz
SRDSS: Self-Reported Disability Status Scale

Edited by C Lovis; submitted 17.03.22; peer-reviewed by U Scholz, HJ Dai; comments to author 07.09.22; accepted 02.10.22; published 10.11.22.

Please cite as:
Chiavi D, Haag C, Chan A, Kamm CP, Sieber C, Stanikić M, Rodgers S, Pot C, Kesselring J, Salmen A, Rapold I, Calabrese P, Manjaly ZM, Gobbi C, Zecca C, Walther S, Stegmayer K, Hoepner R, Puhan M, von Wyl V
The Real-World Experiences of Persons With Multiple Sclerosis During the First COVID-19 Lockdown: Application of Natural Language Processing
JMIR Med Inform 2022;10(11):e37945
URL: https://medinform.jmir.org/2022/11/e37945
doi:10.2196/37945
PMID:36252126

Original Paper

Shared Interoperable Clinical Decision Support Service for Drug-Allergy Interaction Checks: Implementation Study

Sungwon Jung[1], MSc; Sungchul Bae[2], PhD; Donghyeong Seong[1], PhD; Ock Hee Oh[3], PhD; Yoomi Kim[4], PhD; Byoung-Kee Yi[5], PhD

[1]Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea

[2]Data Science Research Institute, Samsung Medical Center, Seoul, Republic of Korea

[3]FirstDIS Ltd, Seoul, Republic of Korea

[4]Electronic Medical Records System Certification Criteria Development Department, Korea Health Information Service, Seoul, Republic of Korea

[5]Department of Artificial Intelligent Convergence, Kangwon National University, Gangwon-do, Republic of Korea

Corresponding Author:
Byoung-Kee Yi, PhD
Department of Artificial Intelligent Convergence
Kangwon National University
1 Ganwondaehakgil, Chuncheon-si
Gangwon-do, 24341
Republic of Korea
Phone: 82 33 250 7672
Email: byoungkeeyi@gmail.com

Abstract

Background: Clinical decision support (CDS) can improve health care with respect to the quality of care, patient safety, efficiency, and effectiveness. Establishing a CDS system in a health care setting remains a challenge. A few hospitals have used self-developed in-house CDS systems or commercial CDS solutions. Since these in-house CDS systems tend to be tightly coupled with a specific electronic health record system, the functionality and knowledge base are not easily shareable. A shared interoperable CDS system facilitates the sharing of the knowledge base and extension of CDS services.

Objective: The study focuses on developing and deploying the national CDS service for the drug-allergy interaction (DAI) check for health care providers in Korea that need to introduce the service but lack the budget and expertise.

Methods: To provide the shared interoperable CDS service, we designed and implemented the system based on the CDS Hooks specification and Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) standard. The study describes the CDS development process. The system development went through requirement analysis, design, implementation, and deployment. In particular, the concept architecture was designed based on the CDS Hooks structure. The MedicationRequest and AllergyIntolerance resources were profiled to exchange data using the FHIR standard. The discovery and DAI check application programming interfaces and rule engine were developed.

Results: The CDS service was deployed on G-Cloud, a government cloud service. In March 2021, the CDS service was launched, and 67 health care providers participated in the CDS service. The health care providers participated in the service with 1,008,357 DAI checks for 114,694 patients, of which 33,054 (3.32%) cases resulted in a "warning."

Conclusions: Korea's Ministry of Health and Welfare has been trying to build an HL7 FHIR-based ecosystem in Korea. As one of these efforts, the CDS service initiative has been conducted. To promote the rapid adoption of the HL7 FHIR standard, it is necessary to accelerate practical service development and to appeal to policy makers regarding the benefits of FHIR standardization. With the development of various case-specific implementation guides using the Korea Core implementation guide, the FHIR standards will be distributed nationwide, and more shared interoperable health care services will be introduced in Korea.

(JMIR Med Inform 2022;10(11):e40338) doi:10.2196/40338

KEYWORDS

clinical decision support; drug-allergy interaction; Health Level 7; Fast Healthcare Interoperability Resources; interoperability; CDS Hooks

Introduction

Clinical decision support (CDS) can improve health care with respect to the quality of care, patient safety, efficiency, and effectiveness [1,2]. In addition, it can reduce the cognitive burden of the physicians upon using the order sets such as procedures and prescriptions [3]. In combination with electronic health records (EHRs), the CDS system influences the behavior of physicians and increases adherence to clinical guidelines [1,4].

However, adopting a CDS system in a health care setting remains a challenge [4,5]. Some hospitals have used in-house developed CDS systems or commercial CDS solutions [6]. Since an in-house CDS system tends to be tightly coupled with a specific EHR system, the functionalities and knowledge base are not easily sharable. On the other hand, a commercial CDS requires costly integration with existing EHR systems both in terms of time and effort. The situations are worse with small- to medium-sized hospitals, including clinics. Lack of budget and expertise prevents them from implementing CDS services [7,8].

Shared interoperable CDS services that enable sharing the knowledge base and expansion of the CDS service can mitigate the previous problems. The services can be implemented using the CDS Hooks, that is, Health Level Seven (HL7) International–published specifications for CDS [9]. The CDS Hooks provides a way to call external CDS services remotely within a provider's workflow [10]. It also uses the HL7 Fast Healthcare Interoperability Resources (FHIR) as a data model. By using the FHIR, the CDS services can provide interoperability to health care providers: tertiary hospitals, small- to medium-sized hospitals, and clinics operating on heterogeneous EHR systems. The result of the decision support is to return the cards displaying text, suggestions, or links to launch a Substitutable Medical Applications, Reusable Technologies (SMART) application [11-14].

Since 2011, the Health Insurance Review and Assessment Service (HIRA) in Korea has provided the drug utilization review (DUR) program as a CDS system containing real-time drug safety data for doctors and pharmacists. The DUR program presents 11 review items, including drug-drug interactions, duplicate prescriptions, and drug regimen dose and duration. The DUR system has been distributed among over 99.8% of health care providers as of 2019 [15-18]. Nonetheless, the adoption of other available CDS services remains a challenge.

Korea's Ministry of Health and Welfare (MoHW) oversees several national initiatives to apply and distribute interoperable health IT standards. As one of several national initiatives, feasibility studies are ongoing to embrace the HL7 FHIR standards [19], widely adopted in the global health care industry [20].

In this study, we focus on developing and deploying the sharable and interoperable CDS service for the drug-allergy interaction (DAI) check based on the CDS Hooks specification at the national level. The main objective of CDS service in the initial stage is technical feasibility and service availability. The DUR program in Korea does not cover the DAI check due to low awareness of the social burden and its prevention for the DAI when setting the review items in 2011 [21]. Global concerns regarding DAI are increasing, and inappropriate medication prescriptions frequently occur in all health care settings [22,23]. Implementation of CDS service for the DAI check is relatively more accessible than other CDS services [12]. The HL7 FHIR standard and CDS Hooks specification allow the CDS service to be sharable, interoperable, and scalable. The study is expected to be a starting point for the national adoption of the HL7 FHIR standards.

Methods

Overview

We developed a shared interoperable CDS system based on CDS Hooks for a DAI check to provide a service to health care providers. The system is triggered by medication orders in the EHR system. When it is evoked, the system checks the DAI and returns recommendations back to the provider. We developed the system in the following steps: (1) requirement analysis, (2) design, (3) implementation, and (4) deployment. In the first step, we identified data elements used for DAI check and classified them into mandatory and optional. We also selected the FHIR resources for contextual information available within an EHR system. Second, we designed concept architecture and web service end points, representational state transfer (RESTful) application programming interfaces (APIs), based on the CDS Hooks structure. We profiled FHIR resources according to data elements and specified the card, a form that represented a result of decision support. We designed a rule engine including a four-step drug-allergy screening logic and knowledge base. Lastly, we implemented components and functions, and deployed the CDS system on a government-managed cloud service called G-Cloud.

The CDS service can simultaneously be used by multiple health care providers, such as tertiary hospitals, small- to medium-sized hospitals, and clinics with their own EHR systems. Health care providers can DAI check using a remote CDS service call when ordering medications. An EHR system creates a request payload with patients' prescriptions and allergy data, and transmits it to the CDS service. The CDS service executes the DAI check logic using the request payload and then returns the result to the EHR system.

Concept Architecture

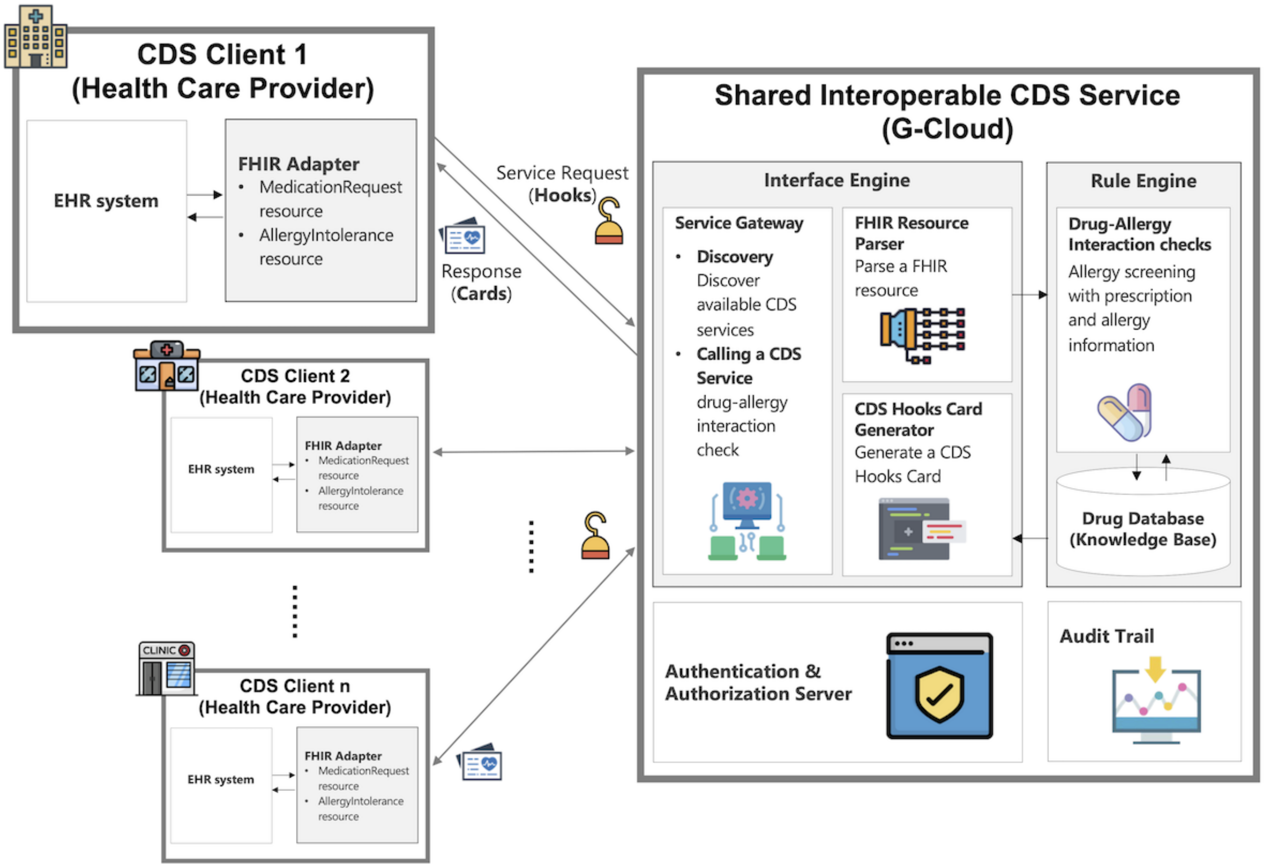
We designed a concept architecture according to the CDS Hooks structure, which consists of CDS services, CDS clients, and cards, as shown in Figure 1. The CDS clients that are EHR systems in health care providers invoke the CDS service through a hook that is an event trigger, and the CDS service provides recommendations using a card to the CDS clients. The CDS system was implemented in version 1.0 of the CDS Hooks specification.

The CDS service was designed as a cloud service and consisted of an interface engine, rule engine, authentication and authorization server, and audit trail. The interface engine has three components: the service gateway, the FHIR resource parser, and the CDS Hooks card generator. The service gateway provides a discovery end point and DAI check end point, and the FHIR resource parser parses request payload data to relay to the rule engine. The CDS Hooks card generator creates decision support results as a card to return to the CDS client. The rule engine checks the DAI using the prescription and patients' allergy information and then returns a result of allergy screening to the CDS Hooks card generator. The authenticate and authorization server authenticates the EHR system using an issued token, and the audit trail monitors which health care providers invoke the service and when and how often they use it.

We applied the CDS Hooks security model with some variations to the CDS service. The CDS Hooks specification provides a security model, such as mutual identification, transport layer security protocol, and JSON web token. We developed the authentication and authorization server to provide a token to CDS clients. The token issued by the CDS service authenticates the CDS client. It reduces the burden on the health care provider's authentication server development and helps a wider adoption of the CDS service. In addition, a whitelist of health care providers is managed based on our risk management strategy.

The CDS client creates an HTTP request to the CDS Hooks service with parameters that include required fields (hook, hookInstance, and context) and optional fields (fhirServer, fhirAuthorization, and prefetch). The context and prefetch fields have the FHIR resources, which are translated by the FHIR adapter. The FHIR adapter was considered instead of an FHIR server since the adoption of the FHIR standard is in its infancy in Korea.

Figure 1. The concept architecture for the shared interoperable CDS system is based on CDS Hooks anatomy. Multiple health care providers simultaneously invoke the shared interoperable CDS service deployed on G-Cloud using a hook and receive a card as a response. CDS: clinical decision support; EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources.


CDS Client 1
(Health Care Provider)
EHR system
FHIR Adapter
• MedicationRequest resource
• AllergyIntolerance resource
Service Request
(Hooks)
Response
(Cards)
CDS Client 2
(Health Care Provider)
EHR system
FHIR Adapter
• MedicationRequest resource
• AllergyIntolerance resource
CDS Client n
(Health Care Provider)
EHR system
FHIR Adapter
• MedicationRequest resource
• AllergyIntolerance resource
CLINIC
Shared Interoperable CDS Service
(G-Cloud)
Interface Engine
Service Gateway
• Discovery
Discover available CDS services
• Calling a CDS Service
drug-allergy interaction check
FHIR Resource Parser
Parse a FHIR resource
CDS Hooks Card Generator
Generate a CDS Hooks Card
Rule Engine
Drug-Allergy Interaction checks
Allergy screening with prescription and allergy information
Drug Database
(Knowledge Base)
Authentication & Authorization Server
Audit Trail

FHIR Resources Profile

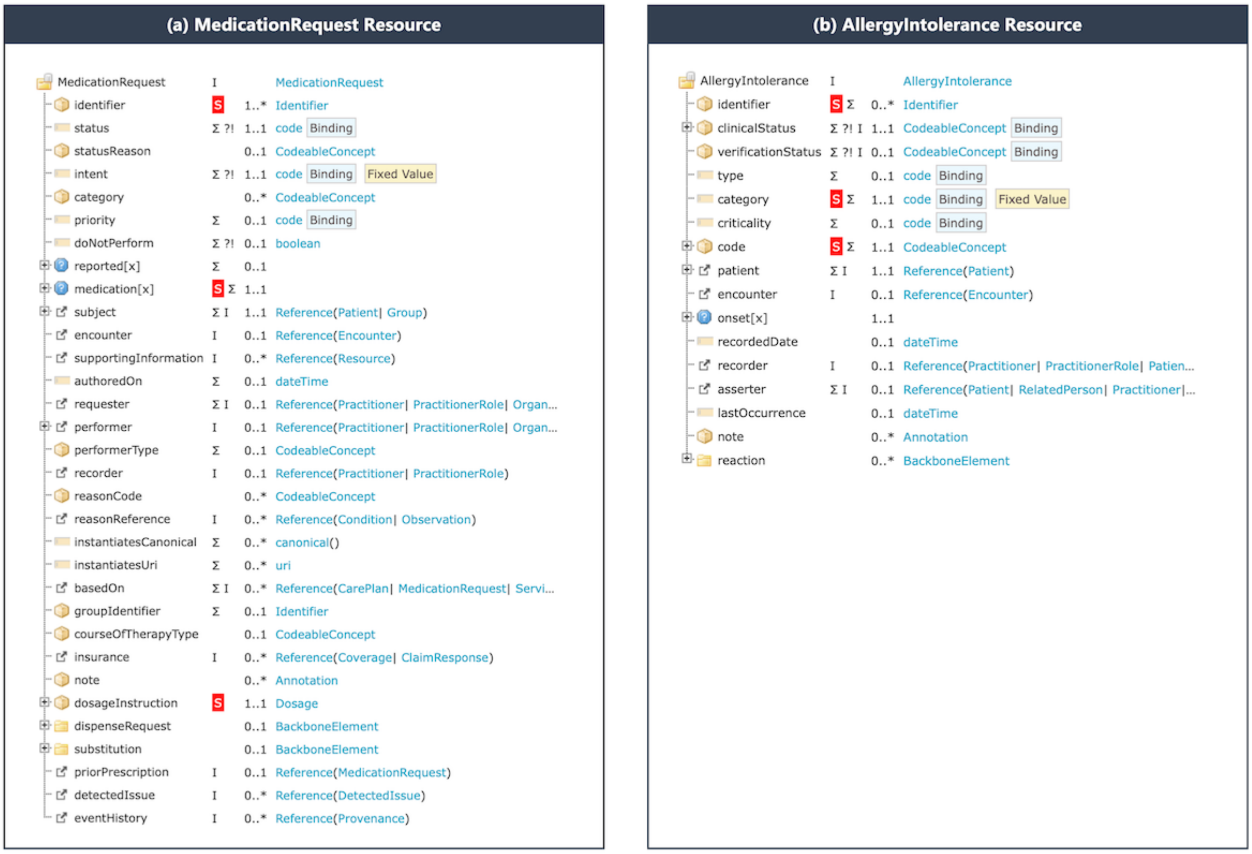
We identified data elements for the DAI check and profiled two FHIR resources, MedicationRequest and AllergyIntolerance, based on the FHIR R4 (v4.0.1) [24] and Korea (KR) Core implementation guide (IG) v1.0.0-STU 1 [25]. The MedicationRequest resource represents a supply of the medication and administration instructions, as shown in Figure 2A [24]. There are two options for representing medication information in the MedicationRequest resource: referencing the Medication resource to the medicationReference element and assigning the medication code directly to the medicationCodeableConcept element. In this profile, we applied the latter because health care providers do not manage the Medication resource. The medicationCodeableConcept element is bound to the Korea Drug (KD) code, the national code system to identify and manage drug products [26]. The cardinality and must support constraints of the MedicationRequest resource are inherited by the KR Core MedicationRequest Profile. The CDS service uses the medication element but not the identifier nor the dosageInstruction elements, although they are marked as must support. Elements designated as a must support are necessary conditions for the FHIR resource to be exchanged, but consumers of the resource do not necessarily have to use all must support elements in principle.

The AllergyIntolerance resource represents a record of a clinical assessment of an allergy or intolerance, as shown in Figure 2B [24]. The category element with the AllergyIntoleranceCategory value set is assigned the fixed value of “medication.” The code element is bound to a proprietary value set developed by the vendor that provides the rule engine of the CDS service, since there is no national code system that identifies the allergy or intolerance.

The cardinality constraints of the AllergyIntolerance resource are inherited by the KR Core IG. The identifier, category, and code elements are marked as “must support.” The profiled resources are published to SIMPLIFIER.NET, one of the FHIR registries.

Two profiled resources are conformant to the KR Core IG in Figure 3. The KR Core IG, a national-level FHIR IG, such as the US Core [27], UK Core [28], Australian Base [29], and Canadian Baseline [30], is essential in the nationwide adoption of the FHIR standards and in building an ecosystem based on the standards. We expect that specific use case FHIR IG based on the resource profiles proposed in this study will be adopted as a national standard in Korea.

Figure 2. The MedicationRequest and AllergyIntolerance resource profile. The resources profiled for the clinical decision support service are inherited from the Korea Core Implementation Guide 1.0.0. Elements with "must support" are marked with an "S" in the red square.



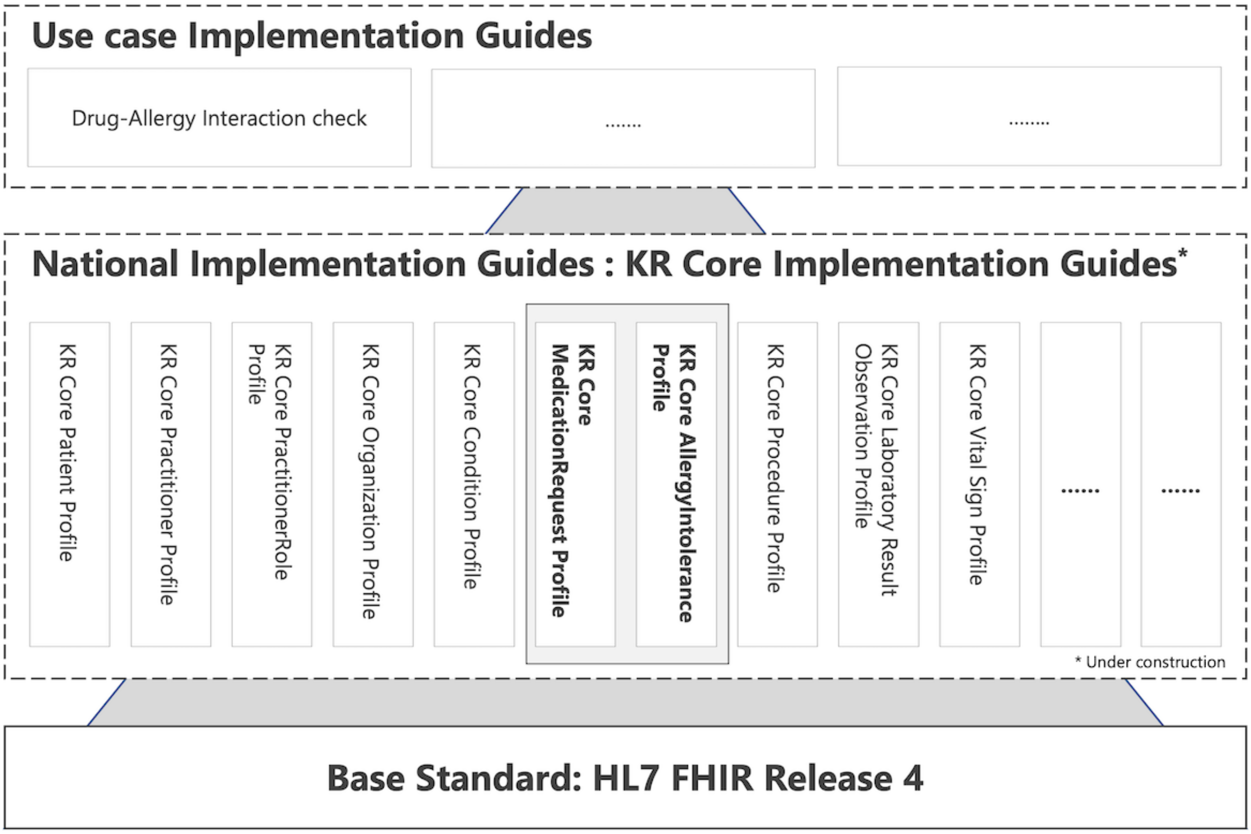
(a) MedicationRequest Resource

MedicationRequest	I		MedicationRequest
identifier	S	1..*	Identifier
status	Σ ?!	1..1	code Binding
statusReason		0..1	CodeableConcept
intent	Σ ?!	1..1	code Binding Fixed Value
category		0..*	CodeableConcept
priority	Σ	0..1	code Binding
doNotPerform	Σ ?!	0..1	boolean
reported[x]	Σ	0..1	
medication[x]	S Σ	1..1	
subject	Σ I	1..1	Reference(Patient| Group)
encounter	I	0..1	Reference(Encounter)
supportingInformation	I	0..*	Reference(Resource)
authoredOn	Σ	0..1	dateTime
requester	Σ I	0..1	Reference(Practitioner| PractitionerRole| Organ...
performer	I	0..1	Reference(Practitioner| PractitionerRole| Organ...
performerType	Σ	0..1	CodeableConcept
recorder	I	0..1	Reference(Practitioner| PractitionerRole)
reasonCode		0..*	CodeableConcept
reasonReference	I	0..*	Reference(Condition| Observation)
instantiatesCanonical	Σ	0..*	canonical()
instantiatesUri	Σ	0..*	uri
basedOn	Σ I	0..*	Reference(CarePlan| MedicationRequest| Servi...
groupIdentifier	Σ	0..1	Identifier
courseOfTherapyType		0..1	CodeableConcept
insurance	I	0..*	Reference(Coverage| ClaimResponse)
note		0..*	Annotation
dosageInstruction	S	1..1	Dosage
dispenseRequest		0..1	BackboneElement
substitution		0..1	BackboneElement
priorPrescription	I	0..1	Reference(MedicationRequest)
detectedIssue	I	0..*	Reference(DetectedIssue)
eventHistory	I	0..*	Reference(Provenance)

(b) AllergyIntolerance Resource

AllergyIntolerance	I		AllergyIntolerance
identifier	S Σ	0..*	Identifier
clinicalStatus	Σ ?! I	1..1	CodeableConcept Binding
verificationStatus	Σ ?! I	0..1	CodeableConcept Binding
type	Σ	0..1	code Binding
category	S Σ	1..1	code Binding Fixed Value
criticality	Σ	0..1	code Binding
code	S Σ	1..1	CodeableConcept
patient	Σ I	1..1	Reference(Patient)
encounter	I	0..1	Reference(Encounter)
onset[x]		1..1	
recordedDate		0..1	dateTime
recorder	I	0..1	Reference(Practitioner| PractitionerRole| Patien...
asserter	Σ I	0..1	Reference(Patient| RelatedPerson| Practitioner|...
lastOccurrence		0..1	dateTime
note		0..*	Annotation
reaction		0..*	BackboneElement

Figure 3. The MedicationRequest and AllergyIntolerance resources profiled through the shared interoperable clinical decision support system are conformed with the KR Core MedicationRequest profile and KR Core AllergyIntolerance profile. FHIR: Fast Healthcare Interoperability; HL7: Health Level Seven; KR: Korea.

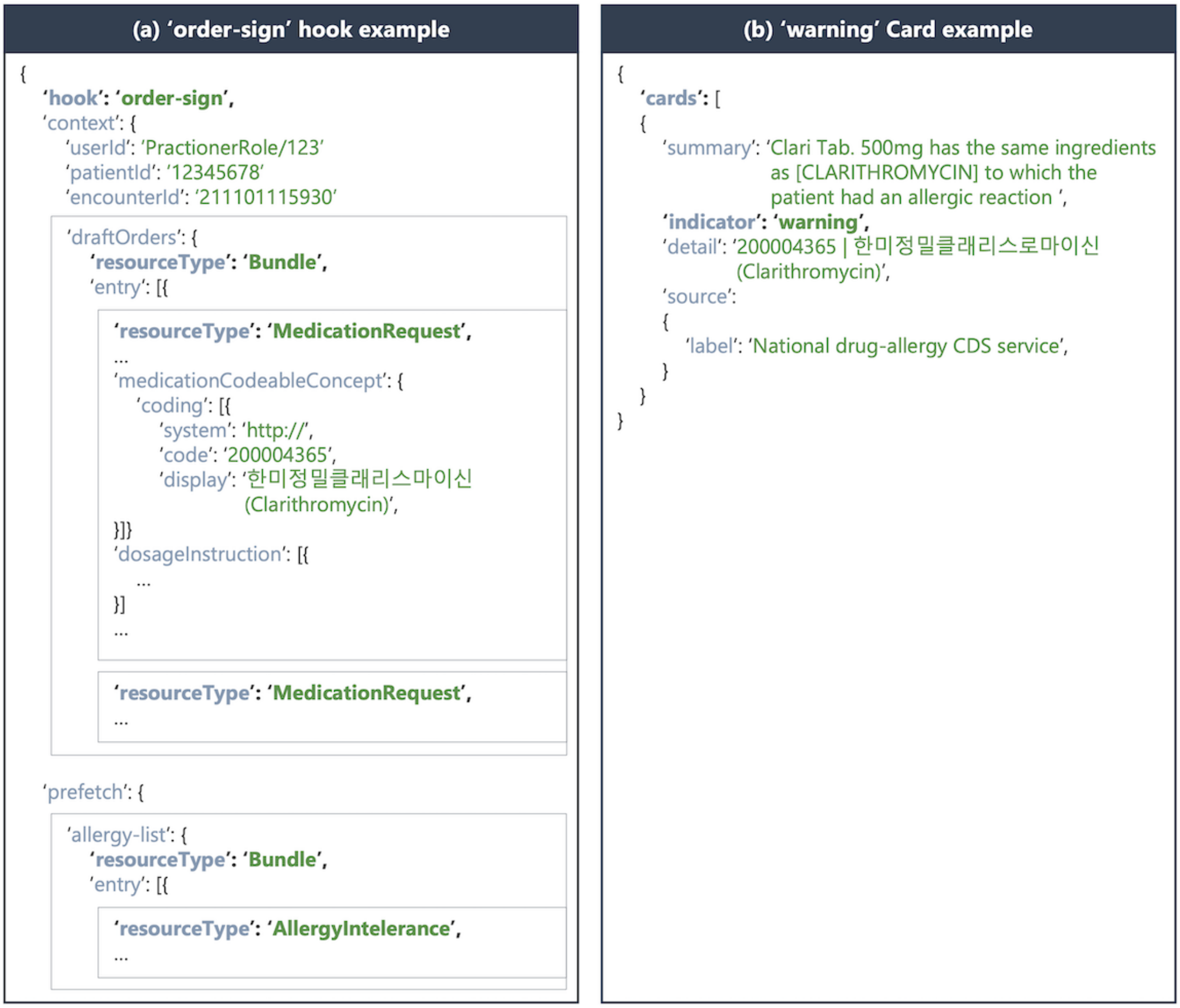

Use case Implementation Guides
Drug-Allergy Interaction check
....
....
National Implementation Guides : KR Core Implementation Guides*
KR Core Patient Profile
KR Core Practitioner Profile
KR Core PractitionerRole Profile
KR Core Organization Profile
KR Core Condition Profile
KR Core MedicationRequest Profile
KR Core AllergyIntolerance Profile
KR Core Procedure Profile
KR Core Laboratory Result Observation Profile
KR Core Vital Sign Profile
......
......
* Under construction
Base Standard: HL7 FHIR Release 4

CDS Service Interfaces and Cards

Two end points were designed and implemented: the discovery and DAI check APIs. The discovery API provides the list of CDS services, including a description of the CDS service and any requested data to be prefetched [9]. The DAI check API is the CDS service using the “order-sign” hook, as shown in Figure 4A. The order-sign hook occurs when the provider is ready to sign one or more orders for a patient, and it has the userId, patientId, and draftOrders as required fields and encounterId as optional. The userId field is included since it is required for the order-sign hook and not used for any other purposes. The CDS service does not distinguish individual providers invoking the service since it does not require a physician ID for DAI checks. The draftOrders field has a Bundle resource that lists MedicationRequest resources. The AllergyIntolerance resources are attached in the prefetch field that describes the relevant data required in the CDS service.

The CDS service responds to the CDS client with cards containing information, suggested actions, and links to launch an application. The DAI check API returns a card with a “warning” indicator, as shown in Figure 4B. The cards are JSON documents and have several fields, such as summary, indicator, and source field. The summary field is a summary message for display to the provider, and the importance of the card is represented by the following indicators: “info,” “warning,” and “critical.” The source field is a source of information displayed on this card.

Figure 4. Examples of order-sign hook and warning card. The order-sign hook has userId, patientId, and draftOrders as required fields, but userId is not used in the clinical decision support (CDS) service for the drug-allergy interaction (DAI) check. The card, the response of the CDS service, includes the results of the DAI check, suggested actions, and links to the launch app.



(a) 'order-sign' hook example

```
{
  'hook': 'order-sign',
  'context': {
    'userId': 'PractionerRole/123'
    'patientId': '12345678'
    'encounterId': '211101115930'
    'draftOrders': {
      'resourceType': 'Bundle',
      'entry': [{
        'resourceType': 'MedicationRequest',
        ...
        'medicationCodeableConcept': {
          'coding': [{
            'system': 'http://',
            'code': '200004365',
            'display': '한미정밀클래리스마이신
                        (Clarithromycin)',
        }]}
        'dosageInstruction': [{
          ...
        }]
        ...
        'resourceType': 'MedicationRequest',
        ...
  'prefetch': {
    'allergy-list': {
      'resourceType': 'Bundle',
      'entry': [{
        'resourceType': 'AllergyIntelerance',
        ...
```

(b) 'warning' Card example

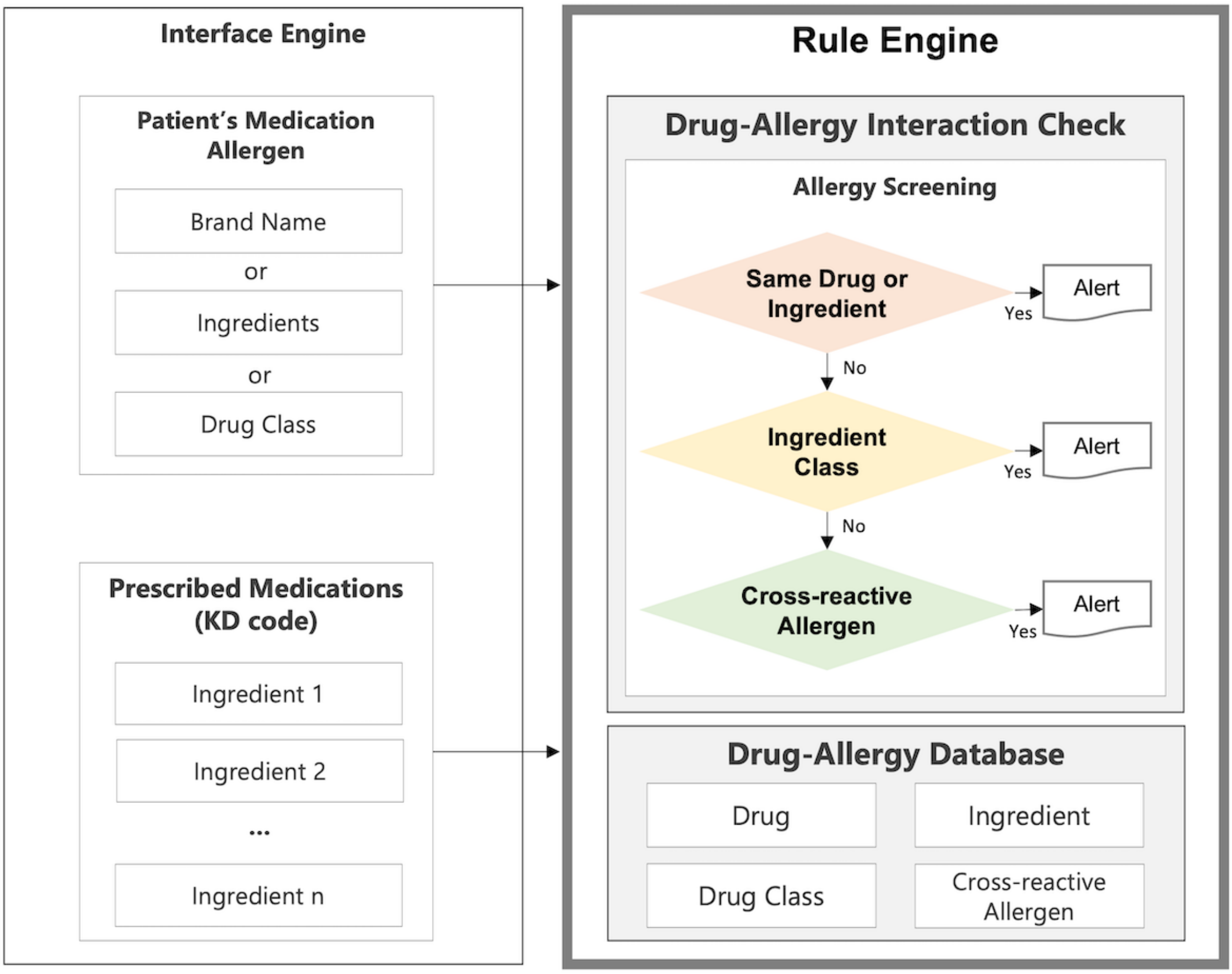
```
{
  'cards': [
  {
    'summary': 'Clari Tab. 500mg has the same ingredients
                as [CLARITHROMYCIN] to which the
                patient had an allergic reaction ',
    'indicator': 'warning',
    'detail': '200004365 | 한미정밀클래리스로마이신
              (Clarithromycin)',
    'source':
    {
      'label': 'National drug-allergy CDS service',
    }
  }
}
```

Rule Engine

We designed and developed the rule engine to check an interaction between a patient's medication allergens and prescribed medications transmitted from a health care provider, as shown in Figure 5. The allergen data from the AllergyIntolerance resource can be a brand name, substance, or drug class. The medication data from the MedicationRequest resource is a brand name coded by the KD code managed by the HIRA. The DAI check is performed in a three-step screening process: (1) check whether allergens and prescribed medications have the same product or ingredient, (2) check whether they belong to the ingredient class, and (3) check whether they have a cross-reactive allergen.

The Drug Allergy database consists of master and relation tables. The master tables are the Drug, Drug Class, Ingredient, and Cross-Reactive Allergen. The Drug and Drug Class tables uniquely identify regulated medicinal products using the KD code as the primary key. The Ingredient table models substances that constitute a medicinal product and includes columns such as ingredient code, name, and synonym. These tables are designed based on the Identification of Medicinal Products, a suite of five International Organization for Standardization standards to facilitate the reliable exchange of medicinal product information. The Ingredient table is related to the Cross-Reactive, Drug, and Drug Class tables by each primary key. In addition to these tables, the drug allergy database has several relation tables used to perform DAI checks.

Figure 5. The three-step drug-allergy interaction check screening process: (1) check whether allergens and prescribed medications have the same product or ingredient, (2) check whether they belong to the same drug or ingredient class, and (3) check whether they have cross-reactive allergens. KD: Korea Drug.



Ethical Considerations

This study did not require ethics approval as no personal data was collected, and no interventions were implemented.

Results

In this study, the national CDS service for the DAI check was developed to ensure the safe use of medicine and was deployed on G-Cloud, a government cloud service established and run by National Computing and Information Service in Korea [31]. The CDS service was launched in March 2021 and has been operated by the Korea Health Information Service. As shown in Table 1, a total of 67 providers participated in the service with 1,008,357 DAI checks for 114,694 patients, of which 33,054 (3.32%) resulted in a "warning" [32]. The results were obtained by analyzing the log data accumulated in the audit trail system.

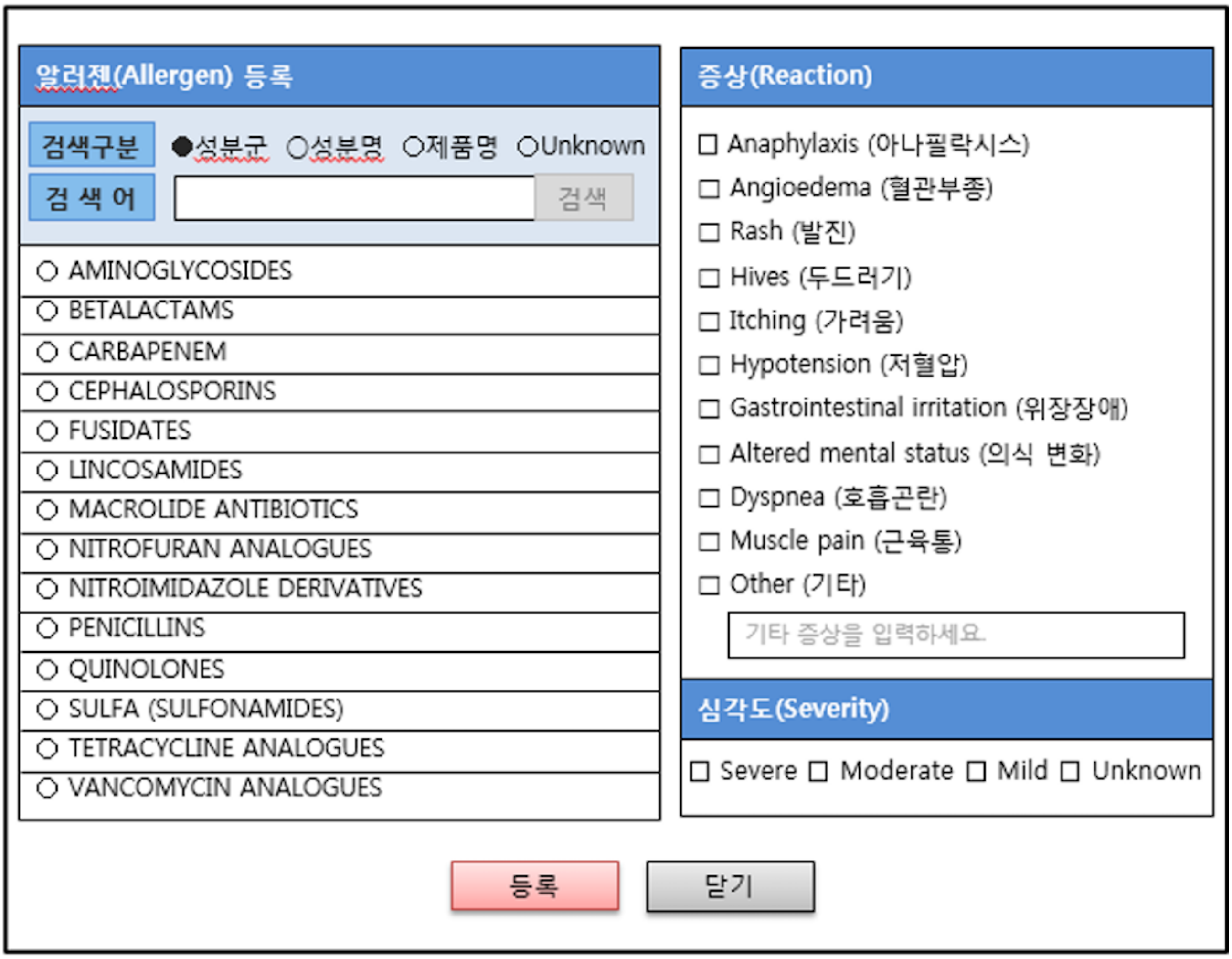
Physicians use the national CDS service for the DAI checks when prescribing medications. The physicians should search the allergen codes provided by the CDS service before calling the CDS service. Since Korea does not yet have a national standard allergy code system, most health care providers store allergy data for the patient as text. To use the CDS service, physicians are also expected to search for an allergy code in the proprietary value set. For this extra step, the CDS service IG provides a reference implementation to inquire about the allergen, allergic reaction, and severity codes, as shown in Figure 6. The health care providers or EHR vendors are expected to develop the component and integrate it with their EHR systems.

Table 1. Results of the shared interoperable CDS service for drug-allergy interaction check in December 2021.

Result category	Amount
Participants, n	
Health care providers	67
Patients	114,694
CDS[a] service requests, n	
Drug-allergy interaction checks	1,008,357
CDS service responses, n (%)	
Warning cards	33,504 (3.32)
No responses	974,853 (96.68)

[a] CDS: clinical decision support.

Figure 6. Screenshot of the reference implementation for a patient's allergen inquiry. To drug-allergy interaction check, physicians should retrieve a patient's allergen code through reference implementation provided by the clincal decision support service.



알러젠(Allergen) 등록

검색구분 ●성분군 ○성분명 ○제품명 ○Unknown

검 색 어 [ ] 검색

○ AMINOGLYCOSIDES
○ BETALACTAMS
○ CARBAPENEM
○ CEPHALOSPORINS
○ FUSIDATES
○ LINCOSAMIDES
○ MACROLIDE ANTIBIOTICS
○ NITROFURAN ANALOGUES
○ NITROIMIDAZOLE DERIVATIVES
○ PENICILLINS
○ QUINOLONES
○ SULFA (SULFONAMIDES)
○ TETRACYCLINE ANALOGUES
○ VANCOMYCIN ANALOGUES

증상(Reaction)

□ Anaphylaxis (아나필락시스)
□ Angioedema (혈관부종)
□ Rash (발진)
□ Hives (두드러기)
□ Itching (가려움)
□ Hypotension (저혈압)
□ Gastrointestinal irritation (위장장애)
□ Altered mental status (의식 변화)
□ Dyspnea (호흡곤란)
□ Muscle pain (근육통)
□ Other (기타)

기타 증상을 입력하세요.

심각도(Severity)

□ Severe □ Moderate □ Mild □ Unknown

Discussion

Principal Results

The study applied the CDS Hooks specification to provide the nationwide shared interoperable CDS service for the DAI check. The CDS service has been deployed on G-Cloud, and all authorized health care providers can use the service simultaneously through RESTful APIs. As of December 2021, 67 health care providers have participated in the initiative. Since the service developed in this study conforms with the CDS Hooks specification, clinical knowledge bases can be shared, and the services can be scalable.

According to the CDS service results report, the rate of warnings that occurred among the CDS service was 7.74% from 29 of the 67 participating hospitals for 1610 patients from May to August 2021. Among the warnings, the most frequent was the cross-reactive allergen check (43.55%), followed by the same drug or ingredient class check (28.77%) and the same product or ingredient check (27.68%). After warning responses from the CDS service, 9.07% of prescriptions were changed, and 90.93% were not changed [33]. Although warning responses occurred from the CDS service, physicians did not change their prescriptions, which had a rate of 90.93%. This proportion is similar to the range of average override alerts, 46.2% to 96.2% [34-36]. To induce physicians to change their prescriptions, additional information and services such as statistical data, research papers, or the SMART application could be provided as evidence.

We designed the CDS system based on serverless FHIR architecture. A CDS service can request additional data regarding the clinical workflow context to the FHIR server at providers via the hook parameters in the CDS Hooks specification. In Korea, the adoption of FHIR standards is in its infancy, and few health care providers have FHIR servers for requesting any additional data. Thus, we applied serverless FHIR architecture, identified the required data in advance, and assigned it in the prefetch field.

As awareness of national allergy codes increases, the MoHW of Korea is developing a national allergy code system. The KR Core AllergyIntolerance profile binds the KR Core AllergyIntolerance Code value set, a renamed version of the AllergyIntolerance Substance value set defined in the FHIR R4. The binding strength of the two value sets is preferred. It is meant to encourage drawing from the specified codes, but it is not required. Currently, there is no national allergy code system available in Korea, and the KR Core AllergyIntolerance Code value set is basically a placeholder for future value set development. Due to the lack of a national allergy code system, we chose to use a proprietary value set. When the national allergy code system is developed, it will replace the value set to draw from the national allergy code system with binding strength required, as well as the KR Core AllergyIntolerance Code value set.

Conclusions

The shared interoperable CDS service for the DAI check based on the CDS Hooks was developed and deployed. The CDS service is currently provided to 67 health care providers. The MoHW has been making efforts to build the HL7 FHIR-based ecosystem in Korea. As one of these efforts, the CDS service initiative was conducted. To promote the rapid adoption of the HL7 FHIR standards, it is necessary to accelerate the practical service development and appeal the benefits of FHIR-based standardization to policy makers; this is the primary purpose of guiding the CDS service. Lastly, with the development of various case-specific IGs based on the KR Core IG, the FHIR standards will be distributed to the health IT industry, and more shared interoperable health care services will be introduced in Korea.

Acknowledgments

This work was supported by the Industrial Strategic Technology Development Program (grant 20020743) funded by the Ministry of Trade, Industry & Energy, Republic of Korea.

Conflicts of Interest

None declared.

References

1. 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]
2. Clinical decision support. HealthIT.gov. URL: https://www.healthit.gov/topic/safety/clinical-decision-support [accessed 2022-10-22]
3. Mills S. Electronic health records and use of clinical decision support. Crit Care Nurs Clin North Am 2019 Jun;31(2):125-131. [doi: 10.1016/j.cnc.2019.02.006] [Medline: 31047087]
4. Marcial L, Blumenfeld B, Harle C, Jing X, Keller M, Lee V, et al. Barriers, facilitators, and potential solutions to advancing interoperable clinical decision support: multi-stakeholder consensus recommendations for the opioid use case. AMIA Annu Symp Proc 2019;2019:637-646 [FREE Full text] [Medline: 32308858]
5. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 04;320(21):2199-2200. [doi: 10.1001/jama.2018.17163] [Medline: 30398550]
6. Cho I, Kim J, Kim JH, Kim HY, Kim Y. Design and implementation of a standards-based interoperable clinical decision support architecture in the context of the Korean EHR. Int J Med Inform 2010 Sep;79(9):611-622. [doi: 10.1016/j.ijmedinf.2010.06.002] [Medline: 20620098]
7. Lee NK, Lee JO. A study on the architecture of cloud hospital information system for small and medium sized hospitals. The Journal of Society for e-Business Studies 2015 Aug 31;20(3):89-112. [doi: 10.7838/jsebs.2015.20.3.089]
8. Seo I. The next healthcarepolicies from the medical perspective. HIRA Policy Brief 2012 Nov 30;1(2):121-126. [doi: 10.52937/hira.21.1.2.121]
9. Current (draft). CDS Hooks. URL: https://cds-hooks.org/specification/current/ [accessed 2022-10-22]
10. CDS Hooks. eCQI Resource Center. URL: https://ecqi.healthit.gov/tool/cds-hooks [accessed 2022-10-21]

11. de Bruin JS, Rappelsberger A, Adlassnig K, Gawrylkowicz J. Exploring methods of implementing Arden Syntax for CDS Hooks. Stud Health Technol Inform 2020 Jun 23;271:191-198. [doi: 10.3233/SHTI200096] [Medline: 32578563]
12. Dolin RH, Boxwala A, Shalaby J. A pharmacogenomics clinical decision support service based on FHIR and CDS Hooks. Methods Inf Med 2018 Dec;57(S 02):e115-e123 [FREE Full text] [doi: 10.1055/s-0038-1676466] [Medline: 30605914]
13. Clinical reasoning. Health Level Seven International. URL: https://www.hl7.org/fhir/clinicalreasoning-module.html [accessed 2022-10-24]
14. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
15. Lee J, Noh Y, Lee S. Evaluation of preventable adverse drug reactions by implementation of the nationwide network of prospective drug utilization review program in Korea. PLoS One 2018;13(4):e0195434 [FREE Full text] [doi: 10.1371/journal.pone.0195434] [Medline: 29641617]
16. Jung S, Jang EJ, Choi S, Im SG, Kim D, Cho S, et al. Effect of a nationwide real-time drug utilization review system on duplicated nonsteroidal antiinflammatory drug prescriptions in Korea. Arthritis Care Res (Hoboken) 2020 Oct;72(10):1374-1382. [doi: 10.1002/acr.24054] [Medline: 31421035]
17. Yang J, Kim M, Park Y, Lee E, Jung CY, Kim S. The effect of the introduction of a nationwide DUR system where local DUR systems are operating--The Korean experience. Int J Med Inform 2015 Nov;84(11):912-919. [doi: 10.1016/j.ijmedinf.2015.08.007] [Medline: 26363001]
18. Kim SJ, Han K, Kang H, Park E. Toward safer prescribing: evaluation of a prospective drug utilization review system on inappropriate prescriptions, prescribing patterns, and adverse drug events and related health expenditure in South Korea. Public Health 2018 Oct;163:128-136. [doi: 10.1016/j.puhe.2018.06.009] [Medline: 30145461]
19. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
20. Healthcare information standardization. Korea Health Information Service. URL: https://www.k-his.or.kr/menu.es?mid=a20203040000 [accessed 2022-10-23]
21. Park J. Social burden of drug allergy and its prevention. Korean J Med 2014;87(6):647. [doi: 10.3904/kjm.2014.87.6.647]
22. Scott IA, Pillans PI, Barras M, Morris C. Using EMR-enabled computerized decision support systems to reduce prescribing of potentially inappropriate medications: a narrative review. Ther Adv Drug Saf 2018 Sep;9(9):559-573 [FREE Full text] [doi: 10.1177/2042098618784809] [Medline: 30181862]
23. Légat L, Van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. Clinical decision support systems for drug allergy checking: systematic review. J Med Internet Res 2018 Sep 07;20(9):e258 [FREE Full text] [doi: 10.2196/jmir.8206] [Medline: 30194058]
24. Welcome to FHIR®. Health Level Seven International. URL: http://hl7.org/fhir/ [accessed 2022-10-12]
25. KR Core Implementation Guide (IG). HINS. URL: https://hins.or.kr/nrc_fhir/index.html [accessed 2022-10-21]
26. Korea Pharmaceutical Information Service. Health Research Review and Assessment Service. URL: https://www.hira.or.kr/eng/news/01/__icsFiles/afieldfile/2013/10/15/brochure_7.KPIS.pdf [accessed 2022-10-21]
27. US Core Implementation Guide. Health Level Seven International. URL: https://www.hl7.org/fhir/us/core/ [accessed 2022-10-21]
28. UK Core Implementation Guide 0.2.0 - STU1. Simplifier.net. URL: https://simplifier.net/guide/UKCoreImplementationGuide0.2.0STU1/Home [accessed 2022-10-21]
29. AU Base Implementation Guide. FHIR CI-Build. URL: http://build.fhir.org/ig/hl7au/au-fhir-base/ [accessed 2022-10-21]
30. Canadian baseline. FHIR CI-Build. URL: https://build.fhir.org/ig/HL7-Canada/ca-baseline/ [accessed 2022-10-21]
31. National Information Resources Service. URL: https://www.nirs.go.kr/eng/index.jsp [accessed 2022-10-07]
32. K-CDS. Standard EMR Framework. URL: https://emrcert.mohw.go.kr/menu.es?mid=a11204030000 [accessed 2022-10-15]
33. Performance evaluation for electronic medical records (EMR) standardization support initiative. Korea Health Information Service. URL: https://tinyurl.com/mpnvn7v6 [accessed 2022-10-23]
34. Poly TN, Islam MM, Yang H, Li YJ. Appropriateness of overridden alerts in computerized physician order entry: systematic review. JMIR Med Inform 2020 Jul 20;8(7):e15653 [FREE Full text] [doi: 10.2196/15653] [Medline: 32706721]
35. Van De Sijpe G, Quintens C, Walgraeve K, Van Laer E, Penny J, De Vlieger G, et al. Overall performance of a drug-drug interaction clinical decision support system: quantitative evaluation and end-user survey. BMC Med Inform Decis Mak 2022 Feb 22;22(1):48 [FREE Full text] [doi: 10.1186/s12911-022-01783-z] [Medline: 35193547]
36. Nanji KC, Seger DL, Slight SP, Amato MG, Beeler PE, Her QL, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc 2018 May 01;25(5):476-481 [FREE Full text] [doi: 10.1093/jamia/ocx115] [Medline: 29092059]

Abbreviations

API: application programming interface

CDS: clinical decision support
DAI: drug-allergy interaction
DUR: drug utilization review
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HIRA: Health Insurance Review and Assessment Service
HL7: Health Level Seven
IG: implementation guide
KD: Korea Drug
KR: Korea
MoHW: Ministry of Health and Welfare
RESTful: representational state transfer
SMART: Substitutable Medical Applications, Reusable Technologies

Edited by C Lovis; submitted 27.06.22; peer-reviewed by J Walsh, J Mandel, ER Khalilian, T Zhang; comments to author 20.07.22; revised version received 16.09.22; accepted 22.10.22; published 10.11.22.

Please cite as:
Jung S, Bae S, Seong D, Oh OH, Kim Y, Yi BK
Shared Interoperable Clinical Decision Support Service for Drug-Allergy Interaction Checks: Implementation Study
JMIR Med Inform 2022;10(11):e40338
URL: https://medinform.jmir.org/2022/11/e40338
doi:10.2196/40338
PMID:36355401

Original Paper

Classifying Comments on Social Media Related to Living Kidney Donation: Machine Learning Training and Validation Study

Mohsen Asghari[1*], PhD; Joshua Nielsen[2*], BSc; Monica Gentili[2*], PhD; Naoru Koizumi[3*], PhD; Adel Elmaghraby[1*], PhD

[1]Department of Computer Science and Engineering, University of Louisville, Louisville, KY, United States

[2]Department of Industrial Engineering, University of Louisville, Louisville, KY, United States

[3]Schar School of Policy and Government, George Mason University, Washington, DC, United States

[*]all authors contributed equally

Corresponding Author:
Mohsen Asghari, PhD
Department of Computer Science and Engineering
University of Louisville
222 Eastern Parkway
Louisville, KY, 40292
United States
Phone: 1 502 403 4483
Email: mohsen.asghari@gmail.com

Abstract

Background: Living kidney donation currently constitutes approximately a quarter of all kidney donations. There exist barriers that preclude prospective donors from donating, such as medical ineligibility and costs associated with donation. A better understanding of perceptions of and barriers to living donation could facilitate the development of effective policies, education opportunities, and outreach strategies and may lead to an increased number of living kidney donations. Prior research focused predominantly on perceptions and barriers among a small subset of individuals who had prior exposure to the donation process. The viewpoints of the general public have rarely been represented in prior research.

Objective: The current study designed a web-scraping method and machine learning algorithms for collecting and classifying comments from a variety of online sources. The resultant data set was made available in the public domain to facilitate further investigation of this topic.

Methods: We collected comments using Python-based web-scraping tools from the New York Times, YouTube, Twitter, and Reddit. We developed a set of guidelines for the creation of training data and manual classification of comments as either related to living organ donation or not. We then classified the remaining comments using deep learning.

Results: A total of 203,219 unique comments were collected from the above sources. The deep neural network model had 84% accuracy in testing data. Further validation of predictions found an actual accuracy of 63%. The final database contained 11,027 comments classified as being related to living kidney donation.

Conclusions: The current study lays the groundwork for more comprehensive analyses of perceptions, myths, and feelings about living kidney donation. Web-scraping and machine learning classifiers are effective methods to collect and examine opinions held by the general public on living kidney donation.

(JMIR Med Inform 2022;10(11):e37884) doi:10.2196/37884

KEYWORDS

living kidney donation; kidney donation; kidney transplantation; text mining; web scraping; NLP; deep learning; neural network; barriers to kidney donation; barriers; awareness; perception; machine learning; online source; online comments

Introduction

Kidney transplantation is the gold standard treatment for patients with end-stage renal disease (ESRD) [1] and can be much more cost-effective than dialysis [2]. Record numbers of transplants have taken place in recent years, but a shortage of donors persists despite recent increases [3]. Currently, the median wait time for a transplant is about 4 years in the United States, and close to 5000 patients die every year on the transplant wait list [4]. Living-donor kidney transplants generally provide better outcomes than deceased donor transplants but are inaccessible to many patients with ESRD, especially among certain racial and ethnic minorities [5,6], because of the potential burdens on donors. Such burdens can include financial costs related to donation and the risk of future kidney failure and death [7,8]. Over the last 2 decades, the US government has implemented programs that reimburse living donors for donation-related expenditures, such as the cost of traveling, medical costs for recovery and possible complications, and time away from the workplace. These programs are, however, known to have had little to no effect on the number of living kidney donors thus far [9].

Several studies have used qualitative approaches to identify possible barriers to kidney donation. These studies have identified several factors that can contribute to decision-making for both living and deceased donation, including the social influence of health care professionals (HCPs) [10], family members [11], and recipients and potential donors [12,13], as well as medical [14] and financial [15,16] barriers. Other factors are related to beliefs and concepts, such as unknown future needs [17] (ie, “What if my family member needs a donation someday?”), a desire for bodily integrity and choice, trust or mistrust of the health care system, religious and cultural beliefs, and a lack of information about donation [10]. Many of these studies, however, focus on identifying factors associated with deceased donation.

Additionally, the data have generally been derived from small samples of interviewees who have already participated in the donation process or from analyses of data from a single transplant center. As such, the extracted data are primarily representative only of those who have had direct experience in living donation. The viewpoints of the general public, who may be curious or have misconceptions about donation but have no direct experience in donation, are thus rarely represented. By leveraging the large volume of opinions and comments available online, this study represents a step toward better understanding of the public’s perception of living donation. At least one other research effort has taken advantage of comments on social media to investigate attitudes about organ donation. Jiang et al [18] found and analyzed 1507 reposts of 141 unique posts related to organ donation on the Chinese microblogging site Weibo; they were able to identify 5 major themes. The authors report that posts on “statistical descriptions” and the “meaningfulness” of organ donation prompted 3 and 2 users, respectively, to express the intention to become an organ donor. Henderson [19] performed a similar analysis.

The specific contribution of this study is the exploration of a machine learning classifier for the collection and analysis of a large database of labeled comments that were written by internet users and collected from multiple public sources. These comments reflect the users’ thoughts, feelings, and concerns regarding living kidney donation (LKD). The authors have made this database available upon request so that researchers on this topic can use the information for further analyses. The current study also examines and discusses the quality of predictions, highlighting particular areas of difficulty with regard to machine classification for further improvement.

Methods

The comments were first collected and processed (the data processing phase). A small portion were then manually classified (annotated and labeled) for use as training data (the annotation phase). The training data were then used to develop a machine learning model that automated the classification process for large volumes of data (the modeling phase).

Data Processing Overview

We created our data set through a process of gathering, filtering, and cleaning data [20]. Data were collected from different sources, including comments on newspaper articles published in the New York Times (NYT) and comments on the social media sites Twitter, YouTube, and Reddit. We manually annotated a small percentage of the data (1174 of 203,219 comments) and designed a neural network to classify the remaining comments. We separated the data set with 2 labels: related or unrelated to LKD. The characteristics of the training and testing data are shown in Table 1.

Figures 1 and 2 illustrate the frequency and distribution of the words in the training and testing data, respectively. The training and testing data were compiled before all the comments were collected, so transfer learning was utilized for the final classification of the Reddit and YouTube comments [21]. The transfer learning model was validated to work sufficiently well on Reddit and YouTube comments by manually inspecting predictions.

Table 1. Characteristics of training and testing data.

Source	Training data (N=934)	Testing data (N=240)
New York Times comments, n (%)	312 (33.5)	83 (34.5)
Tweets, n (%)	622 (66.5)	157 (65.4)
Average words per comment, n	63.2	64.4
Maximum words per comment, n	380	381
Minimum words per comment, n	2	3

Figure 1. Word frequency and distribution for training data.

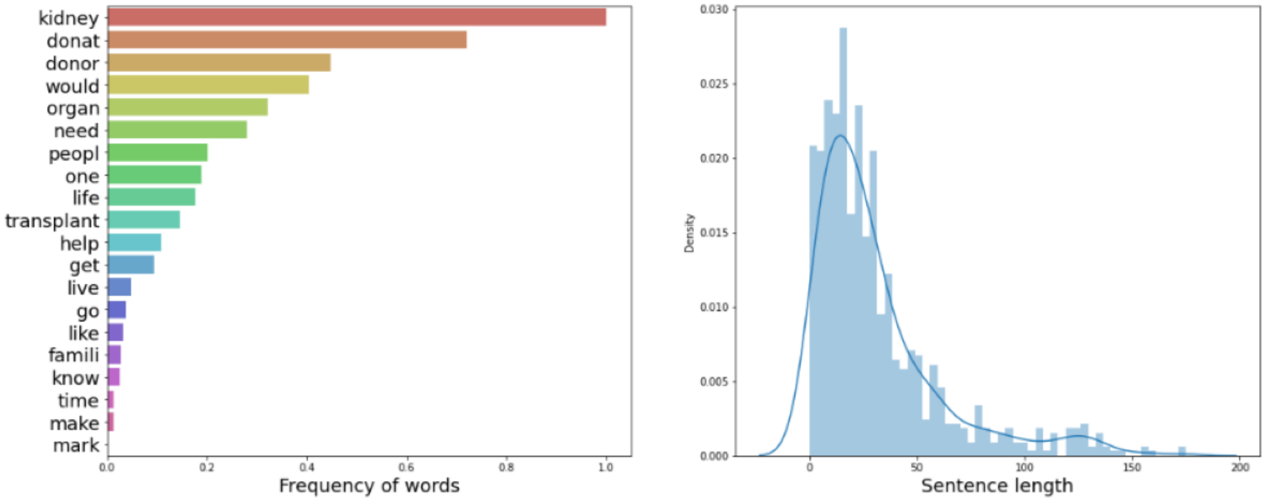
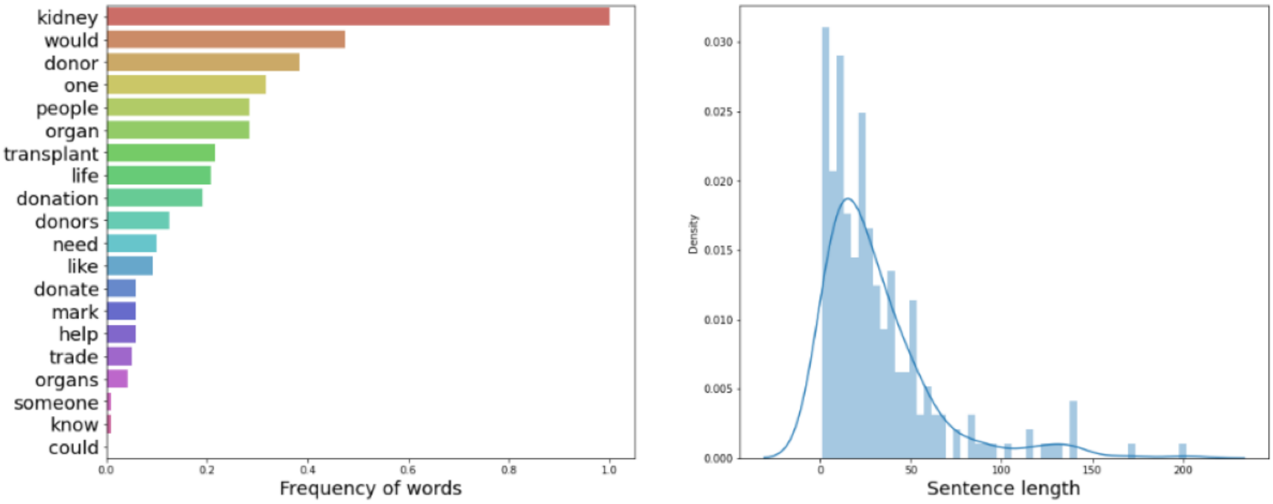


Figure 2. Word frequency and distribution for testing data.



Data Collection

To automate the process of downloading comments from the web, we used the Pushshift.io service for Reddit, Selenium for YouTube, and the application programming interfaces (APIs) from Twitter and the NYT. For each web source, we used search terms aimed at capturing content associated with LKD, while also excluding undesired content (such as political fundraising, which would otherwise appear in searches for terms like “donation”). More details on this exclusion process can be found in Multimedia Appendix 1. Tweets were captured via a live stream over time, but comments from the other 3 sources were captured from any date range allowed by the respective APIs. As YouTube had no API that suited our purposes, we manually searched YouTube for the term “living kidney donation” and identified 17 relevant videos with at least 30 comments each. Table 2 shows how many comments were collected, and over what time period, from each source.

Table 2. Summary of date ranges and numbers of comments (N=203,219).

Source	Date range	Unique comments, n
Twitter	Oct 2020-Apr 2021	148,662
Reddit	Jan 2010-Apr 2021	43,382
New York Times	Jan 2008-Apr 2021	6559
YouTube	Feb 2005-Apr 2021	4616

Class Label Definition

The manual labeling of training data was one of the most important tasks in this study. The purpose of this classification labeling was to determine if a given comment was related to living organ donation. The annotation team worked through 1174 randomly selected comments and determined how each comment should be classified. We assumed at this stage that every comment from every source had equal weight. The process began with 3 annotators collaborating to classify a set of 403 comments, aiming to reach agreement on how the comments should be classified. The remaining 771 comments were classified after the decision criteria were more thoroughly established (the final criteria are described in the following section).

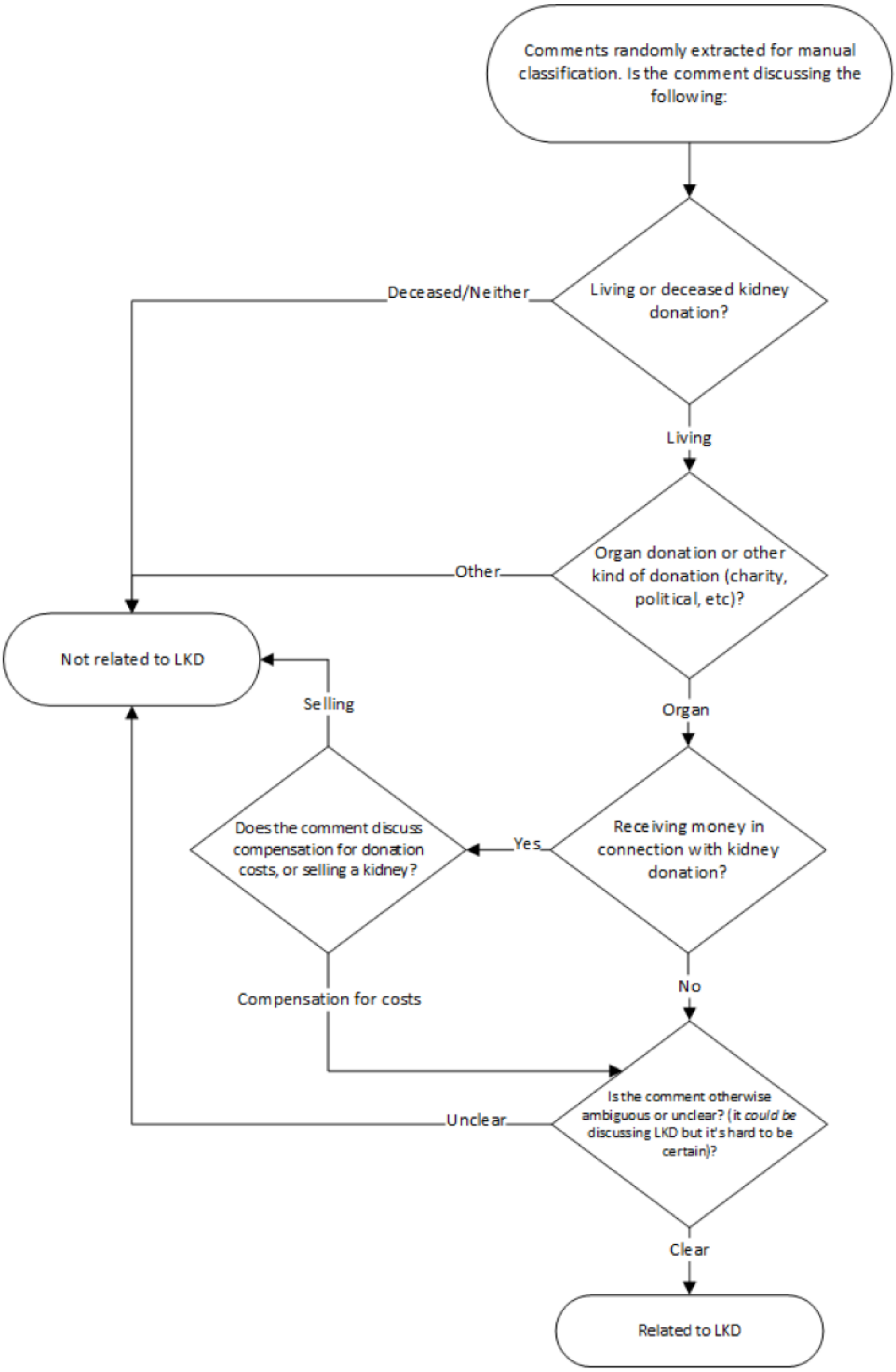
Handling Ambiguity and Other Complexities

Annotations began with a simple idea: capture the comments that mention LKD. But the convoluted reality of human language is rarely simple enough for easy classification, and nuances abound. For example, can we assume a person's sentiments on deceased donation carry over to their opinions on living donation? How should we classify comments in which people express their thoughts on a policy related to LKD even if they do not say whether they personally would donate? To overcome this obstacle, each annotator was given a set of classification criteria to determine whether a comment should be classified as "related."

Even with the explicitly defined classification criteria, the annotation team still encountered significant difficulty in reaching a consensus on many of the comments. During the first stage of annotation, of 403 comments to be annotated, 124 were not classified unanimously. A few guiding principles emerged as the team discussed the dissenting comments. First, while comments explicitly mentioning organ sales and conversations about the illicit organ trade were excluded, the criteria were expanded to allow most other comments that involved cost or finance-related policies about organ donation.

The second principle was to reverse an initial position about encouraging annotators to select "yes" in cases of uncertainty and ambiguity and to instead select "yes" only when they were confident doing so. This last criterion was to clarify that each comment must be viewed as independent from all the other data and that the human annotators should not consider the larger context (ie, other comments in the discussion) or make inferences. This last adjustment represents an important distinction in the way that humans learn compared to the way that machines learn. It is important to note that these criteria forced us to exclude data that could ultimately have been meaningful in order to obtain better performance for the overall model. A flowchart illustrating the decision criteria process is shown in Figure 3. We note that the comments determined to be "not related" were not quantified by the exclusion criteria during the annotation process, so numbers are not available.

Figure 3. Classification criteria for manual labeling of training data. LKD: living kidney donation.


Comments randomly extracted for manual classification. Is the comment discussing the following:
Living or deceased kidney donation?
Deceased/Neither
Living
Organ donation or other kind of donation (charity, political, etc)?
Other
Organ
Not related to LKD
Selling
Receiving money in connection with kidney donation?
Yes
Does the comment discuss compensation for donation costs, or selling a kidney?
No
Compensation for costs
Is the comment otherwise ambiguous or unclear? (it could be discussing LKD but it's hard to be certain)?
Unclear
Clear
Related to LKD

Modeling

We developed a deep neural network to perform automated classification of the remaining comments (Multimedia Appendix 2) using PyTorch 1.11 in Python (version 3.8; Python Software Foundation). The network architecture is shown in Figure 4, with the hyperparameters illustrated by shaded boxes. Table 3 shows possible values for these parameters, each of which was evaluated to determine the best model.

Figure 4. Neural network architecture. NYT: New York Times; RNN: recurrent neural network.

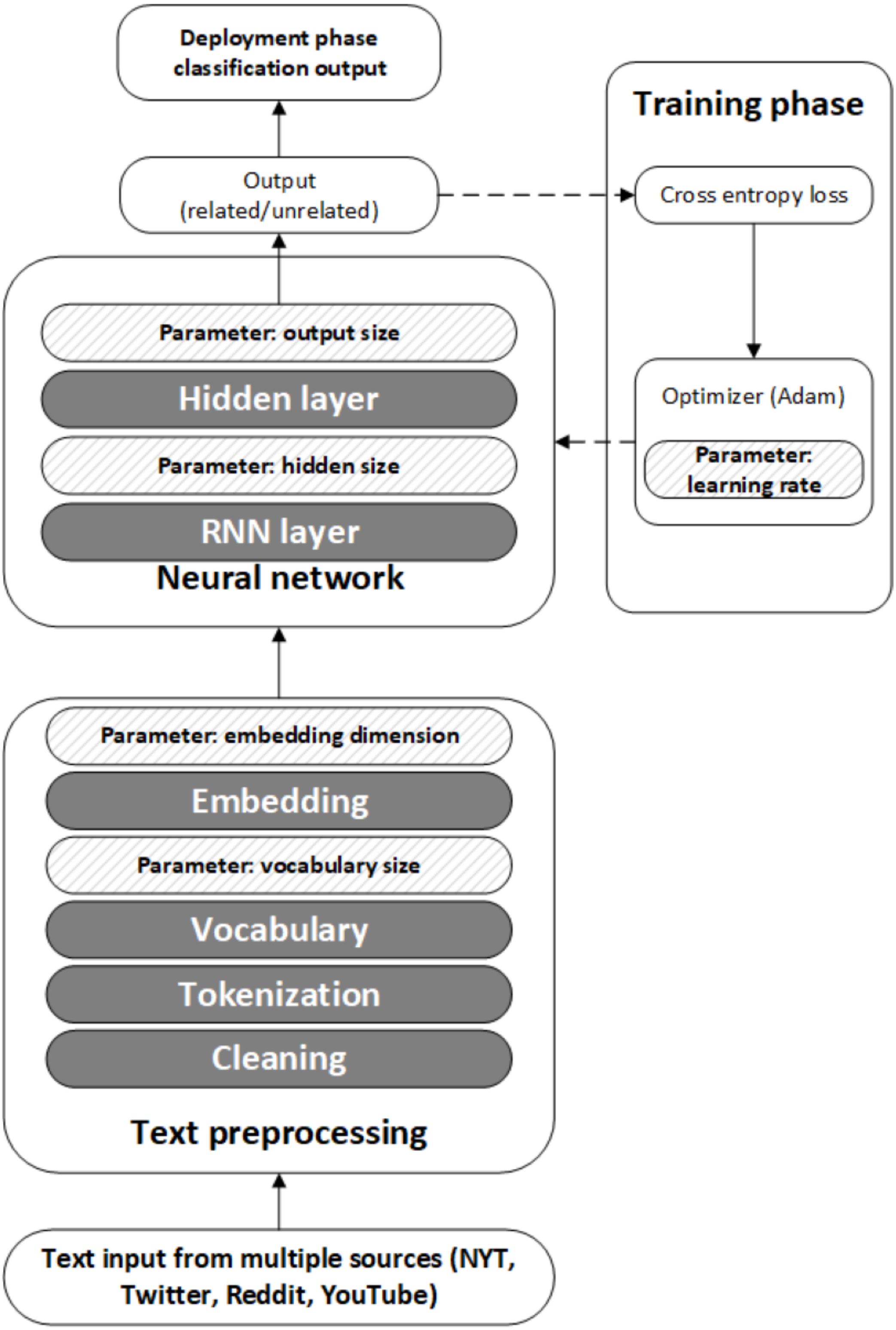


Table 3. Neural network parameters and corresponding experimental values.

Parameter	Range
Tokenization level	Word, character
Embedding layer size	500, 600, 700, 1204, 2048
Hidden layer size	20, 30, 50, 100, 150, 200, 400, 500, 600
Learning rate	0.01, 0.001, 0.0001, 0.00001, 0.000001
Batch size	8, 16, 32, 64, 128

Text Preprocessing

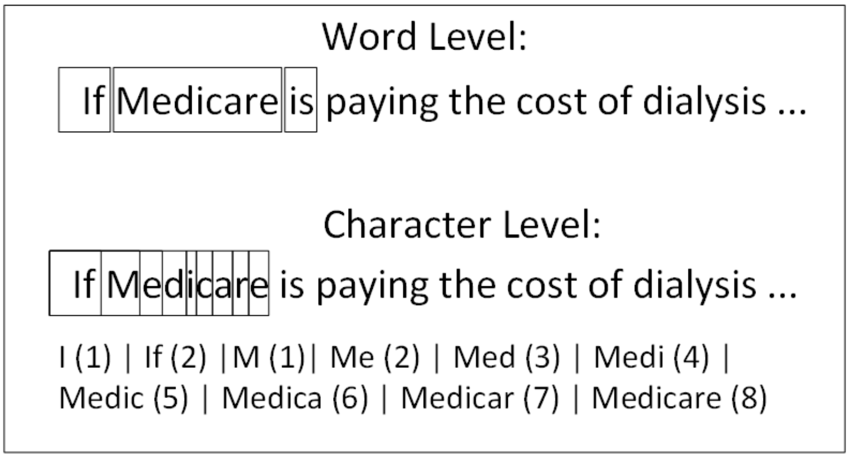
Prior to analyzing the text, documents were cleaned and normalized. The purpose of this text processing was to separate meaningful words from noise. This involved removing strange characters (eg, ¬ and ±), HTML tags, URLs, unnecessary repeated characters (“pleeeease” to “please”), number-character combinations (“401k”), adjusting contractions (“I’ve” to “I have”), and emojis. Words were also stemmed, so that words with the same root but different suffixes (such as “donate,” “donating,” and “donated”) would be treated as the same word (becoming “donat”).

Tokenization was also performed at this stage. Tokenization is the process of separating sentences into smaller parts, such as words and characters. Word level tokenization is a split determined by a space between words, and character level tokenization is the process of dividing a word into different sections based on the length of characters. For example, we created 8 additional tokens from the word “Medicare,” as illustrated in Figure 5.

As the neural network cannot process text, we needed a layer to transform the vocabulary layer to numbers, a process called embedding. There are several techniques for this transformation, such as Google’s Word2Vec [22] and Stanford’s GloVe [23]. We experimented with these tools, but the specific domain of the text topics led to poor performance. To remedy this, we fed our vocabulary (as illustrated in Figure 5) to the Pytorch embedding tool [24], which allows users to train their own embedding layer.

We defined our neural network architecture with 2 layers: a hidden layer (where transformations take place) and an output layer (which determines the final classification). The hidden layer consisted of recurrent neural network nodes that were constructed with a long short-term memory cell [25]. We generated the probability for the output layer such that if the output layer generated a number greater than zero and less than 0.5 for a given comment, it was classified as not related; if it was between 0.5 and 1, it was classified as related. We used CrossEntropy [26] to define the loss function for the training process [27] and used the Adam optimizer [28] to optimize the neural network.

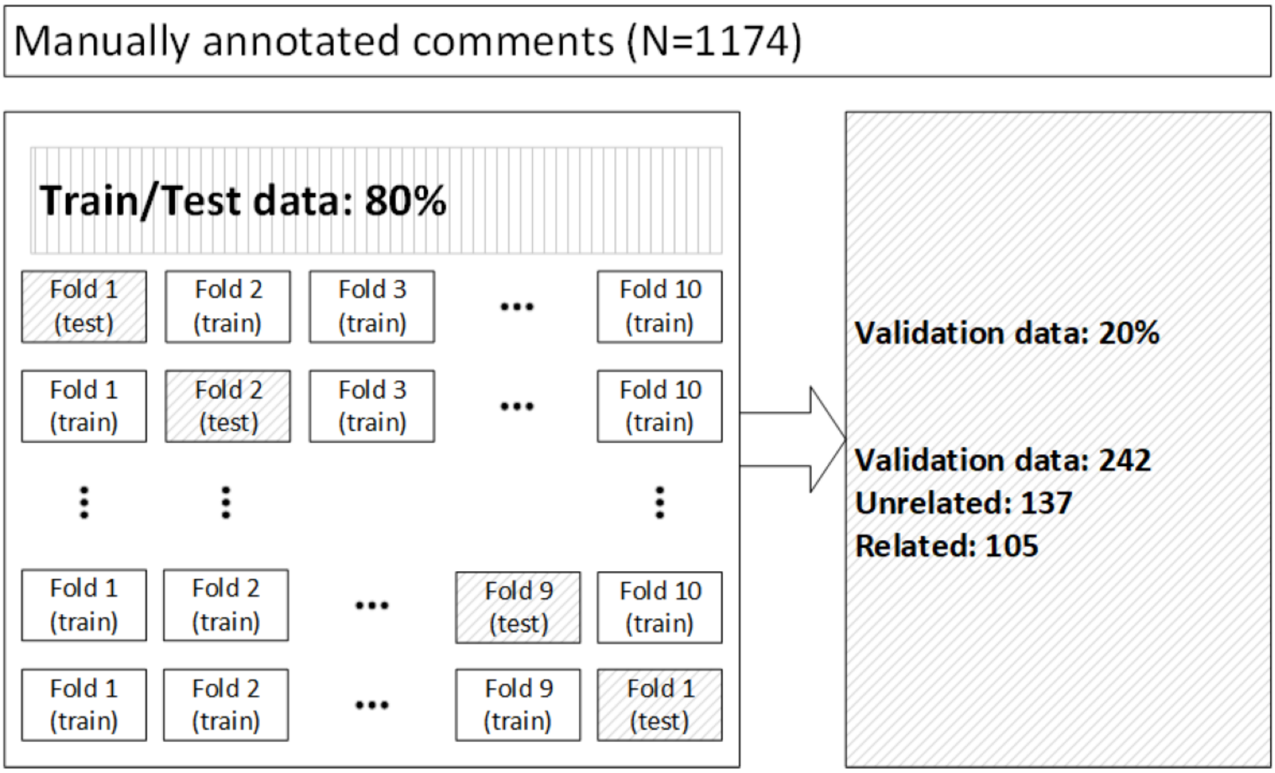
Figure 5. Illustration of word and character tokenization.


Word Level:
If Medicare is paying the cost of dialysis ...
Character Level:
If Medicare is paying the cost of dialysis ...
I (1) | If (2) |M (1)| Me (2) | Med (3) | Medi (4) |
Medic (5) | Medica (6) | Medicar (7) | Medicare (8)

Training and Evaluation Phase

We used a nested K-fold validation procedure to guarantee the necessary model generality [29-31]. In the first iteration, we randomly separated 20% of the data to build the validation data set. The rest of the data (80%) were split into 10 separate folds to be iteratively used as training and testing data. Figure 6 shows the structure of the experiment. We selected K=10 so that we had 10 models to check against our test data set. The purpose of using K-fold cross-validation was to test how well the model could perform on unseen data by training it on small, separate chunks. K-fold validation was also considered useful in comparing the efficacy of word tokenization and character tokenization.

Figure 6. Structure of data training experimentation.



The metrics used to evaluate the performance of the classification model were precision (P), recall (R) and F1 score. The calculation of these metrics is explained in equations 1, 2, and 3, where related comments were treated as positive, and not related comments were treated as negative. The following notation was used: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The precision metric measures how many related comments were correctly classified out of all comments that had been classified as related by the model. On the other hand, the recall metric indicates how many comments were correctly classified out of all the comments that were labeled as related by the annotation process. To select a winning model, the value of both precision and recall should be near 1. The F1 score is the harmonic mean of precision and recall; this measure provides a sense of model generalization. Accuracy (equation 4) is the number of correct classifications out of all classifications made.

Assessment of Machine-Classified Comments

After identifying satisfactory hyperparameters for the model, the model was used to automatically classify the complete data set. To verify the quality of the automated results, a random assortment of 912 comments (219 for the NYT, 222 for Reddit, 187 for Twitter, and 284 for YouTube) for each prediction outcome (ie, “related” and “unrelated”) were read and given an indicator to determine if the classification was correct according to the classification criteria described in the section “Handling Ambiguity and Other Complexities.” False positives (ie, comments incorrectly predicted to be related) were further labeled to identify the error made by the classifier using the categories described in Table 4.

Table 4. Description of false-positive error types.

Classifier error type	Description
Deceased donation	Comment was centered on deceased rather than living kidney donation
Figure of speech	Comment used phrases such as "I'd give a kidney" as a figure of speech or in a joking manner
Insufficient information	Comment had language that was too ambiguous to clearly determine its association with living kidney donation
Irrelevant	Comment was entirely unrelated to living kidney donation (see discussion section for more information)
Kidney stones	Comment mentioned kidney stones with no reference to living kidney donation
Non–living kidney donation policies	Comment expressed opinions on policies related to kidney donation, such as opt-out versus opt-in or legalization of kidney sales, with no information about how such policies might affect the commenter's personal decision regarding donation
Recipient, dialysis, or kidney failure	Comment discussed challenges specifically for (or from the perspective of) a potential kidney recipient, such as kidney failure and dialysis; no information about living kidney donation
Selling or money	Comment discussed the monetary value of a kidney (specifically not used as a figure of speech or joke)

Ethical Approval

The University of Louisville Institutional Review Board provided approval exemption for this study (22.0458).

Results

In this section, we show the quantitative outcomes of our analysis. A testing accuracy of 84% was achieved using the following model hyperparameters: 10-character-gram tokenization, 700 embedding layers, a batch size of 8, 50 hidden layers, and a learning rate of 0.00001. Additionally, precision, recall, and F1 score each achieved 84% in the test data. Once the neural network was trained to achieve the above results, it was used to automatically classify the remaining comments. This yielded 11,027 related comments and 192,192 unrelated comments. Results from further evaluation of the predicted values, as discussed in the section "Assessment of Machine-Classified Comments," are shown in Tables 5 and 6, sorted by comment source. Additional details on the further evaluation can be found in Table S7 in Multimedia Appendix 3.

Table 7 shows the distribution of false positive errors by social media source. We note that many of the irrelevant YouTube comments can be attributed to a single popular video showing an interview with a celebrity whose friend donated a kidney to her.

Table 5. Summary results for sensitivity and specificity of postclassification data.

Sources	False positives (N=576)	True positives (N=336)	False negatives (N=100)	True negatives (N=812)
New York Times, n (%)	107 (18.6)	112 (33.3)	19 (19)	200 (24.6)
Reddit, n (%)	146 (25.3)	76 (22.6)	27 (27)	195 (24)
Twitter, n (%)	159 (27.6)	28 (8.3)	7 (7)	180 (22.2)
YouTube, n (%)	164 (28.5)	120 (35.7)	47 (47)	237 (29.2)

Table 6. Summary results for F1 macro, precision, recall, and accuracy of postclassification data.

Sources	F1 macro (total score of comments was 60.2%), %	Precision (total score of comments was 36.8%), %	Recall (total score of comments was 77.1%), %	Accuracy (total score of comments was 62.9%), %
New York Times	70	51.1	85.5	60.7
Reddit	58	34.2	73.8	47.1
Twitter	46.8	46.8	15	46.8
YouTube	61.2	61.2	42.3	46.2

Table 7. Count of error types by source.

False positives	New York Times (N=107)	Reddit (N=146)	Twitter (N=159)	YouTube (N=164)	Total (N=576)
Deceased donation, n	16	10	0	10	27
Figure of speech, n	0	2	43	3	48
Insufficient information, n	9	39	6	15	69
Irrelevant, n	39	80	60	114	293
Kidney stones, n	0	0	15	0	15
Non–living kidney donation policies, n	25	4	0	2	31
Recipient, dialysis, or kidney failure, n	17	9	23	27	76
Selling or money, n	1	2	12	2	17

Figure 7 shows the confusion matrices for predictions made based on comments from the NYT, Reddit, Twitter, and YouTube, respectively, followed by the confusion matrix for the comments from all sources (in aggregate). We observe that for each source—and overall—the model had greater numbers of false positives than false negatives, illustrating a tendency to overpredict comments as being related.

We observed that 107 of 336 (32.3%) of comments in the related categories were on the topic of personal relationships (Table S7 in Multimedia Appendix 3), which can reasonably be expected, as these are currently the most common type of living donations that take place. We also observed that 293 of 576 (50.1%) of false positives (ie, comments incorrectly predicted to be related) were in the irrelevant category. This category produced the greatest number of false positives from each source. Table 8 shows the other top 2 categories that were most prevalent in misclassifications, along with an example comment to illustrate each.

Figure 7. Confusion matrices for New York Times, Reddit, YouTube, Twitter, and aggregated comments. Clockwise from the top left corner, each quadrant of the confusion matrix shows the true negatives, false positives, true positives, and false negatives. An ideal model will produce quadrants in the top left and bottom right whose color is associated with high values (bright yellow colors), and quadrants in the top right and bottom left whose color is associated with low values (very dark purple colors). NYT: New York Times.

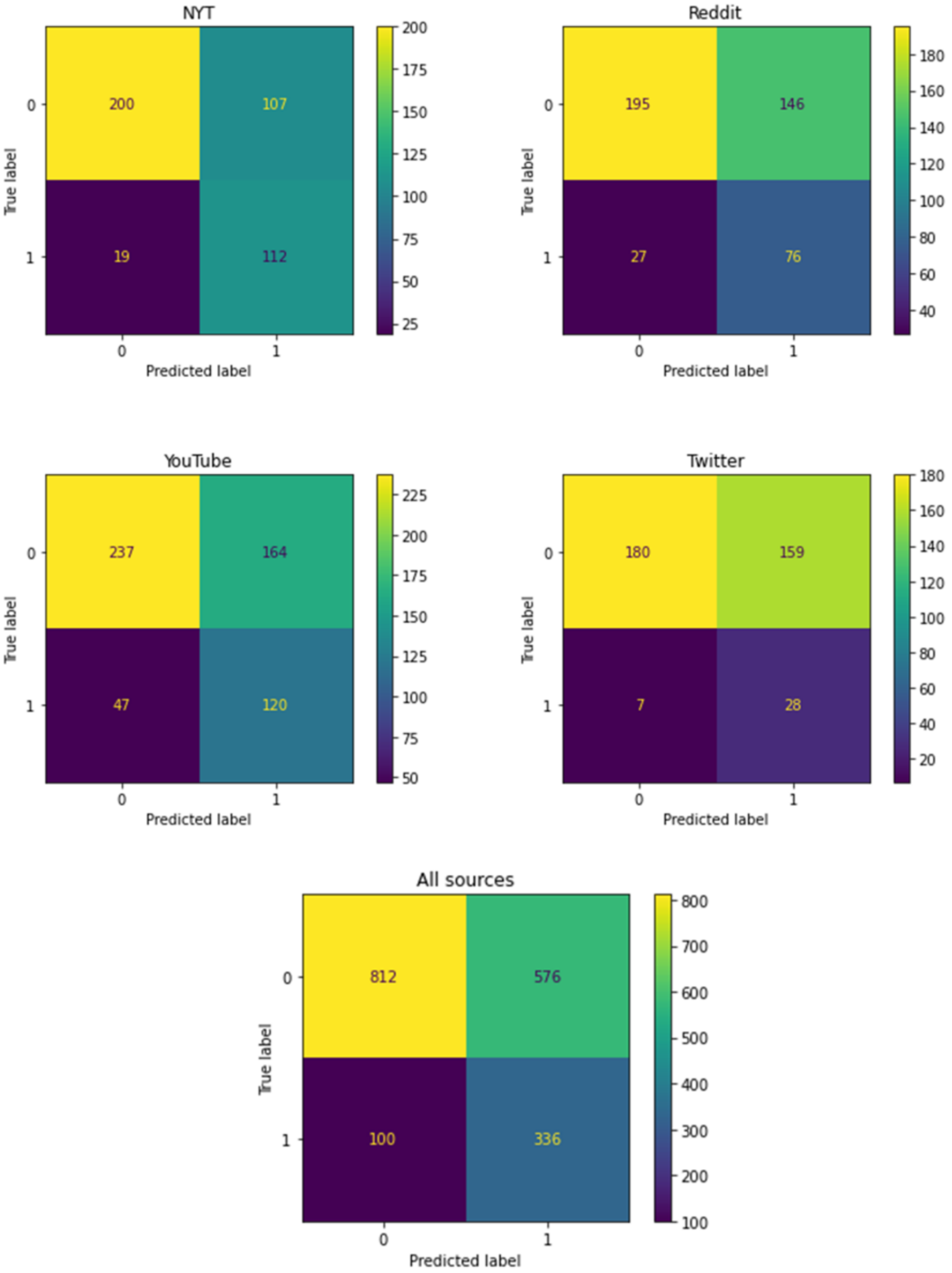


Table 8. Top 2 categories that were most prevalent in their misclassifications, with example comments. Comments are shown “as-is“ after undergoing preprocessing.

Sources/categories	Example comments
New York Times	
Non–living kidney donation policies (25/107 misclassified comments)	“how about making organ donations an opt out process instead of opt in everyone is automatically an organ donator unless they opt out several european countries do this with much success”
Recipient, dialysis, or kidney failure (17/107 misclassified comments)	“my mom was on dialysis for years and died at the age of i was seeing what she went through i would never use dialysis i would get my affairs in order make my peace with god and simply fade away”
Reddit	
Insufficient information (39/146 misclassified comments)	“it really sucks but at that age i wouldn’t even give my grandma one it probably wouldn’t even be recommended”
Deceased donation (10/146 misclassified comments)	“the point is that when you re dead you re dead being on the donor list is the right thing to do no matter what and there is nothing that anyone can say to change that there is no excuse for not being a donor in my eyes”
Twitter	
Figure of speech (43/159 misclassified comments)	“i m going to this even if i have to sell my kidney”
Recipient, dialysis, or kidney failure (23/159 misclassified comments)	“when good things happen to good people my friend s husband finally got a kidney”
Kidney stones (15/159 misclassified comments; this category was unique to Twitter)	"i don t know if it is a kidney stone all i know is it s been days and isn t letting up i thought i maybe pulled a muscle but this isn t muscle pain for sure"
YouTube	
Recipient, dialysis, or kidney failure (27/164 misclassified comments)	“i ve been on dialysis for almost a year I am i m going next week for my evaluation the while process scares me so bad it s so hard but i want it so bad i ll do anything to be normal again”
Insufficient information (15/164 misclassified comments)	“mad respect for this man i would like the courage to do something like this one day”

Discussion

Principal Findings

This study confirmed that the comments available from the internet can provide data on the general perception of living donation. Our trained model identified 11,027 comments related to LKD and 192,192 comments unrelated to LKD. Above, we present a sample distribution of comments that were incorrectly classified and their associated error types. There was a great deal of nuance and subtlety in the comments that could cause confusion for human classifiers, further increasing the difficulty for the machine classifier.

Many users wrote comments expressing their opinions regarding current policies. Though there was disagreement regarding how, nearly all users were supportive of making organs and transplants more accessible. There was notable support for a policy that would give preference or priority to designated or past organ donors when they face the need for organs. In the context of compensation for donation costs, it was also common to observe conversations regarding the legalization of organ sales. The two sides of this were primarily concerns about taking advantage of vulnerable populations and confidence in ethical market self-regulation. The various sources from which the comments were retrieved provided different kinds of comments. Comments that contained opinions about policy were most likely to be retrieved from the NYT, though they were also common on Reddit. There were also several self-reported accounts in the NYT and YouTube comments of someone or their spouse having previously been a living donor.

The character-restricted nature of Twitter meant that comprehensive ideas were less likely to be captured. Twitter was also more likely to produce comments in which people asked for donations or advocated for a loved one in need of an organ. Meaningful comments from YouTube were more often from people who had previous experience with transplants, either as patients or donors. While many of the Reddit comments were of little use, the “ask me anything” (AMA) subreddit provided a veritable treasure trove of information. There were threads written by people who had donated altruistically and invited people to “AMA.” This format, more than any other we encountered, seemed to yield the most thoughtful questions, concerns, and even resolutions to those concerns (to paraphrase one such person upon learning about a voucher system for the donor’s loved ones: “I’ve considered doing this before and never actually [done] anything. This has inspired me to sign up. Thank you!”).

Though there were positive responses from many users, some users were more cynical. One such user expressed the following: “the risk to living donors is also downplayed...people are guilted into acting as living donors only to find themselves at greater risk down the line." Others wrote about frustrating experiences with the medical system or other worries, but we did not observe any blatantly false ideas in the comments. Lack of information was much more common than possession of misinformation.

To efficiently compile relevant information from comments and opinions found on the web, we used deep neural networks trained with specific criteria-driven classification labels. With this approach, we were able to develop a model that could identify comments related to LKD with an expected accuracy of 84%. Though further work remains to refine these results and classify these related comments according to the relevant factors, this first stage of classification indicates that the method could potentially be a valuable tool to extract themes related to barriers to and motivations for living donation. Because the topic is so nuanced, well-defined classification criteria for training data will be a vital part of developing a successful model. It is vital to have multiple people collaborating on training data annotation to ensure uniformity. Without these measures, the viability of this approach becomes less certain.

We note that the sizeable number of comments classified as irrelevant was to be expected to some extent. We suggest the following reasons to explain why our model incorrectly classified irrelevant comments as related to LKD: First, the size of the training data was relatively small compared to the total number of comments classified (1174/203,219). We project that with more (and more correctly labeled) training data, the model would yield better predictions. Second, models based on neural networks tend to have generalization errors that are sometimes identified as gaps [32]. Third, as mentioned above, there exists a great deal of nuance in this topic, and certain words that have no real significance may appear to the model as being important. For example, "parts" could be seen as a word that indicates "parts" of a body (ie, an organ), but it is simply a common word used in many settings.

For deceased kidney donation, there are a handful of studies that have utilized modern computer-science methods to analyze motivations and challenges associated with kidney donation. A recent study [33] discussed the use of natural language processing to glean information about deceased donors and the prospective utility of their kidneys. This information was retrieved from the United Network for Organ Sharing's DonorNet program, in which organ procurement organizations enter raw text about the donors' medical and social history, the history of their admissions, and other noteworthy information. A similar study [34] gathered 342 Spanish articles that contained the text "donacion de organos." The authors found that a positive perception of kidney donation may be a contributing factor to the high rate of kidney donation in Spain. In another study [35], social media posts were collected to study the limitations of social messaging campaigns for deceased kidney donation.

Through the process of manual classification of training data, we observed nearly all barriers noted in the prior literature listed above, as well as early indicators of patterns. For example, the data suggested that the most frequent factors seen in the comments were directly related to the potential impact on prospective donors: considerations of immediate costs and risks of donating and the consequences of such a decision on those close to a donor. Broader influences, such as culture and belief systems, the influence of family members, and perceptions of the medical system, were less relevant to decisions related to living donation and more relevant to decisions related to deceased donation. In our manually labeled data, we did not observe the influence of HCPs as a factor that influenced a prospective donor's decision to donate. Prior research indicates that barriers to donation attributable to HCPs include, for instance, lack of communication between transplant and dialysis teams, lack of training and information among HCPs, and negative attitudes held by some HCPs toward LKD [10].

Our study also recognized that the content and the quality of comments varied rather significantly depending on where they were retrieved. The AMAs of Reddit invited people to ask whatever questions they had, to be answered by someone who had been through the process personally. The downside of this particular resource is that there were only a few AMAs from living kidney donors. Comments from the NYT were more dependent on the content of the article to which they were attached, had no dialogue with the author, and were more conducive to debates on policy than to answering questions from curious prospective donors. Further analysis may provide greater insights into what kinds of internet sources yield the most meaningful information.

Limitations and Future Work

These collected data provide several opportunities for research on LKD. The data can be used for more complicated analysis, such as topic modeling and clustering, with the purpose of detecting barriers and motivations in multisource data sets. Future work may consider the following: instead of a first-stage binary classification, it may be beneficial to consider 4 classifications, such as "irrelevant," "recipient-related," "deceased donation," and "LKD-related." As deceased donation and recipient-related issues are commonly intertwined with conversation about policies, such identification may also help mitigate the misclassification of those topics and reduce the number of entirely irrelevant comments that are erroneously classified as related. Other methods, such as multi-task learning models, could make predictions for comments based on their media source without requiring an independent model for each source.

Additionally, we assumed that each comment should be read independently to aid the model classification. However, it is sometimes possible to maintain an association between comments. For example, in Reddit, each comment has an ID, and if it is a reply, there is a parent ID connecting it to the original comment to which the user is replying. By using this association, the assumption of independence may not be necessary, because it can be better understood that the comment is being written (or not written) in the context of LKD. This would likely help reduce the number of comments which—alone—do not contain enough information to determine their relevance to LKD ("insufficient information").

We observed that there was very little propagation of myths or blatantly false ideas. Among comments that discussed deceased donation (ie, that were unrelated to LKD), there were cynical comments that doctors might reduce life-saving efforts for a dying patient so that an organ could be harvested quickly. While cynicism or frustration with personal experiences appeared in some related comments, misconceptions about LKD were usually nested in expressions of fear or concern (the “risk of donation” category, for example). We suggest that users are more likely to have no (or very little) information about LKD than to have incorrect information. The comments generally indicated that people were curious and prone to ask questions about LKD and wanted to make suggestions about how to increase the number of living donors.

We also acknowledge that more comments could be added to the training data, as the given number of labeled comments was a result of the time-consuming nature of the annotation process. In this exploratory study, we focused on estimating the necessary sample size through a human-annotation process and defining possible labels for the first time. The labeled comments are available upon request from the authors. Finally, we acknowledge that this data is not necessarily representative of all populations. Though internet access continues to expand globally, the distribution of users is not uniform, and each source will have different user bases. For example, according to the 2022 Global Digital Overview Report [36], Reddit users are twice as likely to be men than women, and other studies, discussed in Amaya et al [37], have estimated that between 80% and 90% of global Reddit users are aged 18 to 34 years. Each other source is likely to have its own unique demographic features that should be considered when making inferences from the data.

There is a significant need to understand why people do or do not choose to be living kidney donors. Although prior literature has made contributions toward understanding the context surrounding donation, there is no publicly available data set with information about the thoughts of the broader population on the matter. This project has taken one step toward filling this gap by scraping 203,219 unique internet user comments and tweets and developing a machine-learning classification model to identify comments related to LKD. The documents classified as relevant to LKD were compiled into a single database and are available upon request from the authors. With this database, the groundwork has been laid for more comprehensive analysis of the feelings and ideas that people have surrounding LKD. The data could also be used to identify common misconceptions about donation or information that could lead to changing minds. While rigorous classification of decision-making factors remains to be performed, the findings from this study show that machine learning is a promising tool for the capture and classification of internet comments related to LKD.

Acknowledgments

Funding for this research was provided by an internal grant award from the University of Louisville’s Office of the Executive Vice President for Research and Innovation, the Transdisciplinary/Multidisciplinary Collaboration Research Planning Grant (grant 54152). This study was also partially funded by the National Science Foundation (EAGER: ISN:/1838306) and by the Logistics and Distribution Institute at the University of Louisville.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Additional details on data collection process.
[DOCX File , 13 KB - medinform_v10i11e37884_app1.docx]

Multimedia Appendix 2
Additional information regarding the use of neural network classifiers.
[DOCX File , 14 KB - medinform_v10i11e37884_app2.docx]

Multimedia Appendix 3
Additional information regarding the evaluation of predicted data.
[DOCX File , 21 KB - medinform_v10i11e37884_app3.docx]

References

1. Abecassis M, Bartlett ST, Collins AJ, Davis CL, Delmonico FL, Friedewald JJ, et al. Kidney transplantation as primary therapy for end-stage renal disease: a National Kidney Foundation/Kidney Disease Outcomes Quality Initiative (NKF/KDOQITM) conference. Clin J Am Soc Nephrol 2008 Mar;3(2):471-480 [FREE Full text] [doi: 10.2215/CJN.05021107] [Medline: 18256371]
2. Axelrod DA, Schnitzler MA, Xiao H, Irish W, Tuttle-Newhall E, Chang S, et al. An economic assessment of contemporary kidney transplant practice. Am J Transplant 2018 May;18(5):1168-1176 [FREE Full text] [doi: 10.1111/ajt.14702] [Medline: 29451350]

3. All-time records again set in 2021 for organ transplants, organ donation from deceased donors. Organ Procurement and Transplantation Network. URL: https://optn.transplant.hrsa.gov/news/all-time-records-again-set-in-2021-for-organ-transplants-organ-donation-from-deceased-donors/ [accessed 2022-05-03]
4. Annual Data Report. Scientific Registry of Transplant Recipients. URL: http://srtr.transplant.hrsa.gov/annual_reports/Default.aspx [accessed 2022-10-06]
5. Purnell TS, Hall YN, Boulware LE. Understanding and overcoming barriers to living kidney donation among racial and ethnic minorities in the United States. Adv Chronic Kidney Dis 2012 Jul;19(4):244-251 [FREE Full text] [doi: 10.1053/j.ackd.2012.01.008] [Medline: 22732044]
6. Johansen KL, Chertow GM, Foley RN, Gilbertson DT, Herzog CA, Ishani A, et al. US Renal Data System 2020 Annual Data Report: epidemiology of kidney disease in the United States. Am J Kidney Dis 2021 Apr;77(4 Suppl 1):A7-A8 [FREE Full text] [doi: 10.1053/j.ajkd.2021.01.002] [Medline: 33752804]
7. Muzaale AD, Dagher NN, Montgomery RA, Taranto SE, McBride MA, Segev DL. Estimates of early death, acute liver failure, and long-term mortality among live liver donors. Gastroenterology 2012 Mar;142(2):273-280. [doi: 10.1053/j.gastro.2011.11.015] [Medline: 22108193]
8. Segev DL, Muzaale AD, Caffo BS, Mehta SH, Singer AL, Taranto SE, et al. Perioperative mortality and long-term survival following live kidney donation. JAMA 2010 Mar 10;303(10):959-966. [doi: 10.1001/jama.2010.237] [Medline: 20215610]
9. Chatterjee P, Venkataramani AS, Vijayan A, Wellen JR, Martin EG. The effect of state policies on organ donation and transplantation in the United States. JAMA Intern Med 2015 Aug;175(8):1323-1329. [doi: 10.1001/jamainternmed.2015.2194] [Medline: 26030386]
10. Sandal S, Charlebois K, Fiore JF, Wright DK, Fortin M, Feldman LS, et al. Health professional-identified barriers to living donor kidney transplantation: a qualitative study. Can J Kidney Health Dis 2019;6:2054358119828389 [FREE Full text] [doi: 10.1177/2054358119828389] [Medline: 30792874]
11. Irving MJ, Tong A, Jan S, Cass A, Rose J, Chadban S, et al. Factors that influence the decision to be an organ donor: a systematic review of the qualitative literature. Nephrol Dial Transplant 2012 Jun;27(6):2526-2533. [doi: 10.1093/ndt/gfr683] [Medline: 22193049]
12. Waterman A, Stanley S, Covelli T, Hazel E, Hong B, Brennan D. Living donation decision making: recipients' concerns and educational needs. Prog Transplant 2006 Mar;16(1):17-23. [doi: 10.1177/152692480601600105] [Medline: 16676669]
13. Barnieh L, McLaughlin K, Manns BJ, Klarenbach S, Yilmaz S, Hemmelgarn BR, Alberta Kidney Disease Network. Barriers to living kidney donation identified by eligible candidates with end-stage renal disease. Nephrol Dial Transplant 2011 Mar;26(2):732-738. [doi: 10.1093/ndt/gfq388] [Medline: 20605838]
14. Min K, Koo T, Ryu JH, Jung M, Jongwan H, Jaeseok Y. Barriers to living kidney donation: A single-center experience. Transplantation 2018;102:S504 [FREE Full text] [doi: 10.1097/01.tp.0000543328.29065.bc]
15. Przech S, Garg AX, Arnold JB, Barnieh L, Cuerden MS, Dipchand C, Donor Nephrectomy Outcomes Research (DONOR) Network. Financial costs incurred by living kidney donors: a prospective cohort study. J Am Soc Nephrol 2018 Dec;29(12):2847-2857 [FREE Full text] [doi: 10.1681/ASN.2018040398] [Medline: 30404908]
16. Tushla L, Rudow DL, Milton J, Rodrigue JR, Schold JD, Hays R, American Society of Transplantation. Living-donor kidney transplantation: reducing financial barriers to live kidney donation--recommendations from a consensus conference. Clin J Am Soc Nephrol 2015 Sep 04;10(9):1696-1702 [FREE Full text] [doi: 10.2215/CJN.01000115] [Medline: 26002904]
17. McCormick F, Held PJ, Chertow GM, Peters TG, Roberts JP. Removing disincentives to kidney donation: a quantitative analysis. J Am Soc Nephrol 2019 Aug;30(8):1349-1357 [FREE Full text] [doi: 10.1681/ASN.2019030242] [Medline: 31345987]
18. Jiang X, Jiang W, Cai J, Su Q, Zhou Z, He L, et al. Characterizing media content and effects of organ donation on a social media platform: content analysis. J Med Internet Res 2019 Mar 12;21(3):e13058 [FREE Full text] [doi: 10.2196/13058] [Medline: 30860489]
19. Henderson ML. Social Media in the Identification of Living Kidney Donors: Platforms, Tools, and Strategies. Curr Transplant Rep 2018 Mar;5(1):19-26 [FREE Full text] [Medline: 29805956]
20. Lyko K, Nitzschke M, Ngomo A. New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe. Cham, Switzerland: Springer; 2016:39-61.
21. Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer Learning in Natural Language Processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/n19-5004]
22. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online on September 7, 2013 [FREE Full text] [doi: 10.48550/arXiv.1301.3781]
23. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2014 Presented at: Conference on Empirical Methods in Natural Language Processing; Oct 25-29, 2014; Doha, Qatar p. 1532. [doi: 10.3115/v1/d14-1162]

24. Paszke A, Gross S, Massa F. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. 2019 Presented at: Annual Conference on Neural Information Processing Systems 2019; Dec 8-14, 2019; Vancouver, BC p. 35. [doi: 10.7551/mitpress/11474.003.0014]
25. Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. arXiv. Preprint posted online on November 21, 2016 [FREE Full text] [doi: 10.48550/arXiv.1611.06639]
26. de Boer P, Kroese D, Mannor S, Rubinstein R. A tutorial on the cross-entropy method. Ann Oper Res 2005 Feb;134(1):19-67. [doi: 10.1007/s10479-005-5724-z]
27. Zhang Z, Sabuncu MR. Generalized cross entropy loss for training deep neural networks with noisy labels. 2018 Presented at: 32nd Conference on Neural Information Processing Systems; Dec 2-8, 2018; Montreal, QC.
28. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv. Preprint posted online on January 30, 2017 [FREE Full text] [doi: 10.48550/arXiv.1412.6980]
29. Cawley G, Talbot N. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:107 [FREE Full text]
30. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 2014 Mar 29;6(1):10 [FREE Full text] [doi: 10.1186/1758-2946-6-10] [Medline: 24678909]
31. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. 1995 Presented at: 14th International Joint Conference on Artificial intelligence; Aug 20-25, 1995; Montreal, QC p. 45.
32. Kantardzic M, Aly AA, Elmaghraby AS. Visualization of neural-network gaps based on error analysis. IEEE Trans Neural Netw 1999;10(2):419-426. [doi: 10.1109/72.750572] [Medline: 18252539]
33. Placona AM, Martinez C, McGehee H, Carrico B, Klassen DK, Stewart D. Can donor narratives yield insights? A natural language processing proof of concept to facilitate kidney allocation. Am J Transplant 2020 Apr;20(4):1095-1104 [FREE Full text] [doi: 10.1111/ajt.15705] [Medline: 31736193]
34. Greco F, Monaco S, Di TM. Emotional text mining and health psychology: the culture of organ donation in Spain. EasyChair. Preprint posted online on August 7, 2019 [FREE Full text]
35. Bail CA. Cultural carrying capacity: Organ donation advocacy, discursive framing, and social media engagement. Soc Sci Med 2016 Sep;165:280-288. [doi: 10.1016/j.socscimed.2016.01.049] [Medline: 26879407]
36. Digital 2022 Global Digital Overview. DataReportal. URL: https://datareportal.com/reports/digital-2022-global-overview-report [accessed 2022-08-17]
37. Amaya A, Bach R, Keusch F, Kreuter F. New data sources in social science research: things to know before working with Reddit data. Soc Sci Comput Rev 2019 Dec 18;39(5):943-960. [doi: 10.1177/0894439319893305]

Abbreviations

AI: artificial intelligence
AMA: ask me anything
API: application programming interface
ESRD: end-stage renal disease
FN: false negative
FP: false positive
HCP: health care professional
LKD: living kidney donation
NYT: New York Times
P: precision
R: recall
TN: true negative
TP: true positive

Edited by T Hao; submitted 10.03.22; peer-reviewed by A Placona, D Pessoa, T Zhang; comments to author 20.04.22; revised version received 18.08.22; accepted 19.09.22; published 08.11.22.

Please cite as:
Asghari M, Nielsen J, Gentili M, Koizumi N, Elmaghraby A
Classifying Comments on Social Media Related to Living Kidney Donation: Machine Learning Training and Validation Study
JMIR Med Inform 2022;10(11):e37884
URL: https://medinform.jmir.org/2022/11/e37884
doi:10.2196/37884
PMID:36346661

Original Paper

Developing an Automated Assessment of In-session Patient Activation for Psychological Therapy: Codevelopment Approach

Sam Malins[1*], PhD; Grazziela Figueredo[2*], PhD; Tahseen Jilani[2*], PhD; Yunfei Long[3*], PhD; Jacob Andrews[4*], PhD; Mat Rawsthorne[5*], BA, CGMA; Cosmin Manolescu[1*], MSc; Jeremie Clos[2*], PhD; Fred Higton[6*], PhD; David Waldram[6*]; Daniel Hunt[7*], PhD; Elvira Perez Vallejos[8*], PhD; Nima Moghaddam[9*], PhD

[1]Specialist Services, Nottinghamshire Healthcare NHS Foundation Trust, Nottingham, United Kingdom
[2]School of Computer Science, University of Nottingham, Nottingham, United Kingdom
[3]School of Computer Science and Electronic Engineering, University of Essex, Essex, United Kingdom
[4]Mindtech Medtech Co-operative, University of Nottingham, Nottingham, United Kingdom
[5]Hilltop Digital Lab Ltd, Stockport, United Kingdom
[6]Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom
[7]School of English, University of Nottingham, Nottingham, United Kingdom
[8]Nottingham Biomedical Research Centre Mental Health and Technology Theme, University of Nottingham, Nottingham, United Kingdom
[9]School of Psychology, University of Lincoln, Lincoln, United Kingdom
[*]all authors contributed equally

Corresponding Author:
Sam Malins, PhD
Specialist Services
Nottinghamshire Healthcare NHS Foundation Trust
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 7811737725
Email: sam.malins@nottingham.ac.uk

Abstract

Background: Patient activation is defined as a patient’s confidence and perceived ability to manage their own health. Patient activation has been a consistent predictor of long-term health and care costs, particularly for people with multiple long-term health conditions. However, there is currently no means of measuring patient activation from what is said in health care consultations. This may be particularly important for psychological therapy because most current methods for evaluating therapy content cannot be used routinely due to time and cost restraints. Natural language processing (NLP) has been used increasingly to classify and evaluate the contents of psychological therapy. This aims to make the routine, systematic evaluation of psychological therapy contents more accessible in terms of time and cost restraints. However, comparatively little attention has been paid to algorithmic trust and interpretability, with few studies in the field involving end users or stakeholders in algorithm development.

Objective: This study applied a responsible design to use NLP in the development of an artificial intelligence model to automate the ratings assigned by a psychological therapy process measure: the consultation interactions coding scheme (CICS). The CICS assesses the level of patient activation observable from turn-by-turn psychological therapy interactions.

Methods: With consent, 128 sessions of remotely delivered cognitive behavioral therapy from 53 participants experiencing multiple physical and mental health problems were anonymously transcribed and rated by trained human CICS coders. Using participatory methodology, a multidisciplinary team proposed candidate language features that they thought would discriminate between high and low patient activation. The team included service-user researchers, psychological therapists, applied linguists, digital research experts, artificial intelligence ethics researchers, and NLP researchers. Identified language features were extracted from the transcripts alongside demographic features, and machine learning was applied using k-nearest neighbors and bagged trees algorithms to assess whether in-session patient activation and interaction types could be accurately classified.

Results: The k-nearest neighbors classifier obtained 73% accuracy (82% precision and 80% recall) in a test data set. The bagged trees classifier obtained 81% accuracy for test data (87% precision and 75% recall) in differentiating between interactions rated high in patient activation and those rated low or neutral.

Conclusions: Coproduced language features identified through a multidisciplinary collaboration can be used to discriminate among psychological therapy session contents based on patient activation among patients experiencing multiple long-term physical and mental health conditions.

(JMIR Med Inform 2022;10(11):e38168) doi:10.2196/38168

KEYWORDS

responsible artificial intelligence; machine learning; cognitive behavioral therapy; multimorbidity; natural language processing; mental health

Introduction

Background

One psychological therapist can vary significantly from another in how effective they are for their patients [1,2]. Furthermore, individual psychological therapists do not necessarily, on average, improve their effectiveness with time or experience [3]. In addition, the beneficial effects of psychological therapies have not grown in many areas, and in some cases, effectiveness has declined over time [4,5]. Given that time and experience alone do not seem to improve effectiveness, there are currently few evidence-based means of helping psychological therapists improve their efficacy. This situation is unhelpful for patients, with significant differences in effectiveness among the psychological therapists they may see. It is also unhelpful for psychological therapists and psychological therapy services with few scalable, cost-effective means of supporting practitioners to improve their effectiveness. There have been calls for systematic, objective, and routine means of measuring the quality of psychological therapy content [6,7], and the application of artificial intelligence (AI) may offer part of the solution, especially in combination with text classification and other natural language processing (NLP) techniques.

AI is defined as a form of technology that (1) is to some degree able to perceive the environment and real-world complexity; (2) collects and interprets information inputs; (3) can perform decision-making, including the ability to learn and reason; and (4) can achieve predetermined goals [8]. Increasingly, AI has been used to categorize and evaluate the contents of psychological therapy sessions in research. In face-to-face psychological therapy, supervised learning models have achieved reliable automation of psychological therapy competency assessments, with particular advances in motivational interviewing and more recently cognitive behavioral therapy [9,10]. In messaging- and internet-based psychological therapy, a bottom-up, unsupervised learning approach has been used to identify the types of language used where clinical improvement is significantly more likely and, conversely, where it is less likely [11,12].

There are several potential benefits to these approaches. First, automated evaluation of psychological therapy could offer scalable, routine assessment of psychological therapy interactions where human coding can be too time consuming and costly [13,14]. Second, AI offers the potential to improve identification and verification of prognostic markers in psychological therapy contents, with associated trainable skills for therapists, which may either be difficult to identify from human coding or where important markers are hard to discover because research of sufficient scale is impractical with human raters. Overall, this approach could offer psychological therapists ongoing feedback on their practice, as routinely recommended [15]. This would allow continual improvements in effectiveness when coupled with, for example, deliberate practice techniques to enhance therapeutic microskills [16,17].

However, none of the current uses of AI in psychological therapy contents have focused on patients experiencing multiple comorbidities (or multimorbidity). This is significant, given that differences among therapists are more pronounced among patients with more complex problems, and patients experiencing multimorbidity generally have poorer prognoses [18]. In addition, more active participation and engagement during health care consultations can have an especially positive effect on long-term physical health, mental health, and service use among patients experiencing multimorbidity [19]. This is particularly important because the majority of treatment and care for multimorbid conditions is undertaken by the patients themselves [20]. Furthermore, the ability of patients in this group to self-manage their care is highly affected by clinician responsiveness and interaction style [21,22]. This suggests that specific in-session process markers may be suitable for automated identification and classification in a patient group where psychological therapy is at greater risk of failure, and interaction style can have an important impact on engagement and prognosis. Current evidence has also been largely restricted to either face-to-face psychological therapy or messaging-based treatment. Less attention has been paid to the large and growing use of videoconferencing psychological therapy since the onset of the COVID-19 pandemic [23].

The important issues of algorithmic trust and participatory approaches to development have also not been sufficiently addressed in current applications of AI to psychological therapy. In recent years, significant concerns have arisen regarding the increasing pervasiveness of algorithms and the impact of automated decision-making in health care, alongside the poverty of research into applying AI systems in practice [24]. This means that AI systems are being developed without sufficient involvement or consideration of stakeholders affected by AI decisions. Particularly problematic is the lack of transparency surrounding the development of these algorithmic systems and their use [25].

Within the field of mental health, the engagement and involvement of key stakeholders, including service users, have been identified and recommended as part of the process of developing trustworthy AI applications [26,27]. Stakeholder engagement is one of the pillars of responsible research and innovation [28] and is central to this study to increase the trustworthiness and relevance of emerging AI applications in psychological therapy. As well as increasing trust in AI, the involvement of stakeholders (including service users) can help address systematic biases in AI systems that can replicate human prejudices in the decisions made [29,30]. At this stage in the nascent use of AI for analyzing psychological therapy content, it may be important to establish methods for using AI responsibly in this particular context [31].

A recently developed psychological therapy rating tool may provide an opportunity to address some of the current gaps in the evidence around the use of AI for psychological therapy evaluation. The consultation interactions coding scheme (CICS) [32] was developed to rate individual psychological therapy interactions, turn by turn, based on patient activation. Patient activation has become a significant, well-used, and well-researched concept in health care, particularly for people experiencing multimorbidity [33,34]. Patient activation is the degree to which a person feels confident and able to be actively involved in managing their own health [35]. Patient activation is distinct from other related motivation and engagement constructs because it more specifically focuses on how engagement and motivation are expressed in consultation interactions between health care users and health care professionals [36]. The patient activation measure (PAM) is the established means of assessing patient activation in research and clinical practice [37]. However, as a retrospective questionnaire, the PAM may not be able to fully inform interventions designed to increase patient activation, which often involve adjusting interaction style during health care consultations [38,39]. Therefore, an assessment of patient activation focused on interactions within consultations could be instructive to health care professionals.

The CICS classifies interactions into themes or interaction types (eg, *action planning*) and assigns a rating to each interaction type based on the level of patient activation. Higher scores denote greater patient activation. Ratings on the CICS have been shown to be associated with working alliance, therapist competence, multiple physical and mental health outcomes, and important clinical changes within therapy among patients with multimorbidity receiving psychological therapy over videoconferencing [32,40,41]. The CICS could address some of the key gaps in AI use for psychological therapy, particularly among patients with multimorbidity and in applications of remote psychological therapy. It may, therefore, offer a basis for an explainable, automated psychological therapy rating tool.

Aims

This study’s aims were as follows:

1. Involve end users and stakeholders in applying participatory elements of an explainable AI methodology to coproduce an initial, automated version of the CICS (autoCICS).
2. Assess the performance of the autoCICS ratings compared with human rating reliability.
3. Identify key language features associated with high and low patient activation as well as different interaction types.

Overall, a participatory methodology, which helps to build trust among stakeholders, was applied to the responsible design and development of an autonomous psychological therapy rating system.

Methods

Data Source

Source data included 128 hours of audio data from remotely delivered cognitive behavioral therapy (rCBT) from 53 participants in a randomized controlled trial of rCBT versus usual care for people with severe health anxiety using urgent care at a high rate [42]. Participants were randomly allocated to rCBT plus usual care (n=79) or usual care alone (n=77). There were 78 participants randomized to rCBT, and 1 participant was randomized to usual care but offered rCBT in error. Their data are included in the analysis. Therefore, the total sample is 79. Participants randomized to rCBT were offered up to 15 sessions of rCBT delivered via videoconferencing software (54/79, 68%) or the telephone (14/79, 18%; the remaining participants—11/79, 14%—did not attend any sessions). Most of the participants were not seeking psychological therapy when recruited (69/79, 87%), and most reported multimorbidity (75/79, 95%).

The randomized controlled trial recruited 156 participants from UK primary and secondary health care settings. Participants were adults (aged ≥18 years) who had received ≥2 unscheduled or urgent consultations with any health care provider in the previous 12 months and were identified as highly anxious about their health. Participants were excluded if they were experiencing an acute medical condition requiring ongoing assessment, but those with comorbid common mental health problems or chronic physical conditions such as depression or chronic pain were intentionally included.

Of the 79 possible participants, 53 (67%) were included, having (1) attended ≥1 rCBT sessions and (2) consented to treatment sessions being recorded and extracts anonymously reported. The structured clinical interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition [43], was completed with participants at baseline, assessing for criteria of mental disorders. Long-term physical health conditions were also recorded from baseline patient interviews.

Four psychological therapists delivered rCBT using an established treatment protocol [42]. Of the 4 therapists, 2 were women, and 2 were men; 2 had doctoral-level clinical psychology training, and 2 had master’s-level psychological therapy training.

Of the 128 included sessions, 98 (76.5%) were first and second sessions, and 30 (23.4%) were identified as sessions of potential clinical importance: occurring directly before a sudden sustained improvement, sudden deterioration, or dropout or were the center session in a series where little or no outcome change occurred. The group of 98 sessions (total 42,064 turns of speech) was used to develop and train the initial model, and the other 30 sessions (total 9,239 turns of speech) were used as a holdout sample to test the model once developed. This split fitted with the separation of early sessions and clinically relevant later sessions available. It also approximated to the established 80:20 percentage split for training and test data sets.

Ethics Approval

Ethics approval was obtained from the National Research Ethics Service, London-Riverside Committee (14/LO/1102).

CICS Categories

The CICS categorizes each in-session turn of speech and rates the level of patient activation. A turn of speech is defined as the words spoken by one party until the other party speaks; when the other speaker begins speaking, the first speaker's turn of speech is deemed to have ended. First, a topic is assigned for the turn of speech from ≥1 of the CICS themes using observable criteria (Textbox 1).

Once an interaction theme is allocated, the level of patient activation present in this interaction is rated. Scores range from +2 for interactions showing observable, high levels of patient activation and engagement to −2 for interactions showing observable indications of low patient activation and disengagement. The CICS rating level allocated is linked to established levels of patient activation (Table 1 presents overall level descriptors for CICS themes and comparator patient activation levels; Table 2 presents an example of level descriptors for the *evaluations of self or therapy* theme). The 2 higher levels of patient activation (3 and 4, equivalent to CICS +1 and +2) are linked with positive health outcomes, and the 2 lower levels (1 and 2, equivalent to CICS −2 and −1) are associated with poorer health outcomes across a range of domains [44]. The CICS coders were trained using a published manual [45].

CICS ratings are defined on the basis of a therapist-patient interaction combined. This aims to address the key issue of responsiveness in psychological therapy. Therapist responsiveness is defined as behavior that is influenced by emerging context, such as a therapist changing their verbal response in line with changes in patient presentation [46]. This kind of responsiveness is an important contributor to therapists' effectiveness [47]. Accounting for this type of responsiveness aims to give therapists feedback on their behavior within specific patient contexts; for example, previous machine learning studies of text-based psychological therapy have identified therapeutic praise (eg, "Well done") from therapists as predictive of better outcomes [11]. However, these therapist utterances must occur in the context of specific patient interactions, which is not accounted for when only the therapist response is considered.

All CICS themes have achieved good-to-excellent interrater reliability (intraclass correlation coefficients=0.60-0.80), and most achieved convergent validity with cognitive behavioral therapy competence and working alliance (r_ss=0.72-0.91). The *problem or context description* interaction theme (rated present or absent) has shown moderate-to-substantial interrater reliability (κ=0.54-0.61) and negative associations with working alliance and therapist competence (r_s=−0.71 and −0.47) [32].

Textbox 1. Description of consultation interactions coding scheme themes.

Interaction theme and description

- Action planning and idea generation: discussion of specific plans or potential plans for activities outside the session
- Evaluations of self or therapy: offering a personal assessment of therapy or of one of the parties in therapy
- Information discussion: giving, receiving, or requesting specific information
- Noticing change or otherwise: where changes are reported that relate to therapeutic work, or a lack of change is described despite efforts to bring it about
- Other: where interactions were not related to therapy; most commonly, these interactions involved resolving technical issues associated with videoconferencing
- Problem analysis and understanding: an analysis or understanding of a problem is given or received
- Problem or context description: description of problems or contexts surrounding problems
- Structuring and task focus: where verbal efforts to structure, plan, or progress the session are offered or sought; conversely, where sessions deviate from any relevant topic without intervention from either party

Table 1. Consultation interactions coding scheme (CICS) scores and equivalent, mapped patient activation levels (adapted from the study by Deeny et al [20]).

CICS level	CICS-level descriptor	Mapped PAM[a] level	PAM-level descriptor and percentage of patients at each level[b]
+2	A high level of patient activation and focus is observable; an interaction usually led by the patient. This would include patient-initiated therapeutic activity not cued or primed by the previous therapist interaction	4	Level descriptor: "I'm my own advocate." Patients who are confident in developing and adopting behaviors and practices to manage their health, such as care planning or self-monitoring. Such individuals may be connecting with supportive others (13% of respondents)
+1	Significant patient activation is observable but with less leadership. Typically, this would be a therapeutically active interaction, led or guided by the therapist, which the patient endorses and develops with their contributions	3	Level descriptor: "I'm part of my health care team." Patients who seem to be taking action, for example, setting goals for their health (such as adhering to a medically advised diet) or collaborating in development of a care plan with health care providers, but may still lack the confidence and skill to maintain these (46% of respondents)
0 or neutral	These are interactions where few or no observable positive or negative interaction features are apparent with regard to patient activation. These interactions are deemed to be neutral—neither beneficial nor detrimental to the outcome. The same code is applied if a theme is absent. This includes interactions where therapists make suggestions or comments with little or no observable sense of how the patient receives them	N/A[c]	N/A
−1	Hypothesized to be therapeutically unhelpful interactions in a minor way. This includes interactions that show the start of unaddressed disagreements or reluctance to engage with therapeutic activities. Low levels of patient activation and involvement are observed	2	Level descriptor: "I could be doing more." Patients who may manage some low-level aspects of their health but struggle in many aspects of their care, such as engaging with care planning (19% of respondents)
−2	Hypothesized to be interactions that would be contradictory to most therapeutic guidance. This would include argumentative or obstructive interactions where the patient and potentially the therapist appear disengaged, unfocused, and oppositional to therapeutic activity	1	Level descriptor: "My clinician is in charge of my health." Patients tend to feel overwhelmed by managing their own health and may not feel able to take an active role in their own care. They may not understand what they can do to manage their health better and may not see the link between healthy behaviors and good management of their condition (22% of respondents)

[a] PAM: patient activation measure.

[b] Data taken from a UK sample of 9348 primary care patients [20].

[c] N/A: not applicable.

Table 2. Level descriptors and exemplar quotes for the evaluations of self or therapy consultation interactions coding scheme theme.

Level	Level descriptor[a]	Exemplar quote
+2	Patient-initiated statements of *self-efficacy,* patient *acknowledgment or pride* at therapeutic achievement, or *positive evaluations* of therapy or the therapist that are initiated by the patient	• Patient: "Like I had a panic attack on Friday so randomly...and I was so good, like I dealt with it so well...I was so good at sort of, like, joking around with myself and I was like yeah, just stay here, just like breathe, like, and I remember thinking, like, I know that, like, no one, these people sitting next to me, literally have no idea because part of me was just like right just carry on because it's going to pass, it's going to pass." • Therapist: "Yeah." [P01078]
+1	Therapist-initiated positive evaluation; as in the previous row, patient agrees with *development and summary orcorrections*	• Therapist: "It sounds like you did exactly the right thing...how you addressed your worry; you know reflecting on it and actually, you know, taking action...rather than just sitting ruminating and going deeper into worry. Sounds like you did the right thing." • Patient: "Yeah. I think reflecting is the best thing I ever did because I was so scared, I was so worried about the outcome...but when I looked at it, it's not my responsibility." [P01108]
0	Therapist-initiated positive evaluation; patient acknowledges with *no development* or very low–level acknowledgment by the patient	• Therapist: "You handled those thoughts well by, you know, not letting them become more catastrophizing by recognizing for what they were and managing to handle them pretty well." • Patient: "Mmm." [P01096]
−1	Therapist's positive evaluations, as in the previous row, are *undermined to some degree* by the patient or *somewhat negatively focused* self-evaluations or statements about therapy or therapist	• Therapist: "Yeah, that's huge. How do you feel about yourself, given that you've done all this stuff this week?" • Patient: "Well, I'm really pleased with this week, but I'm still cross about the things that I didn't do, as opposed to being pleased about the things that I did do." [P03014]
−2	*Self-denigrating or self-critical* statements or a *self-critical focus* on therapeutic tasks that have not been completed to the exclusion of those that have been completed by the patient	• Patient: "I wouldn't say that I have that much control over my way of dealing with things." • Therapist: "Really?" • Patient: "Yes." [P01007]

[a] Italics add emphasis to the key component of the level descriptor.

Focusing on Problem or Context Description

The most reliable finding from predictive modeling with the CICS so far is that the greater the proportion of sessions taken up with *problem or context description* interactions, the poorer the outcome. In this way, *problem or contextdescription* interactions were predictive of poorer generalized anxiety, health anxiety, depression, quality of life, and general health across a 12-month follow-up [41]; they also negatively predicted well-being rated across therapy sessions and significantly reduced in frequency directly before sudden sustained outcome improvements [40]. Despite being associated with poorer outcomes, *problem or context description* interactions are conceptualized as neutral, not negative, interactions—describing problems is a necessary and normal part of psychological therapy; however, excessive focus on problem description alone may crowd out space for other types of interactions, particularly those where higher patient activation is indicated and greater active engagement may be stimulated. Therefore, *problem or contextdescription* interactions are scored present or absent as opposed to higher or lower patient activation as in the case of other interaction themes, with the aggregate score being the percentage of the session rated for the theme.

Given the central importance of *problem or context description* interactions to the prognostic validity of the CICS, we first focused autoCICS classification modeling on identifying *problem or context description* interactions versus other interactions. Second, given the importance of higher patient activation across the other CICS interaction themes, autoCICS classification modeling also focused on identifying interactions categorized as higher versus lower levels of patient activation.

Data Preprocessing

Each session was transcribed verbatim, with any identifying information removed during transcription, and transcripts were then checked for anonymity by the raters. Each transcribed turn of speech was coded in NVivo software (version 12.0; QSR International) by three trained raters using the CICS (SM, CM, and NM). A third pass was carried out in preprocessing to assign a master code to each turn of speech accounting for the previous raters' decisions. Overlapping codes were also removed in the master code because they would not be processed effectively when generating classification models in the autoCICS approach. The two possible positive ratings on the CICS (+1 and +2) were collapsed into a single positive category (1), and the possible neutral and negative ratings (0, −1, and −2) were collapsed into a single negative category (0), sacrificing some granularity in the data to increase data subgroup sizes used to train the predictive models. General demographic features were added as predictor variables alongside language features, including participant age and sex, alongside therapist sex. Features were also added to represent the natural grouping of transcribed speech: speech from the same patient, as well as interactions occurring at the beginning, middle, or end of a session (dividing the total turns of speech into three). Minimal demographic features were used with the aim of both addressing common end-user concerns about data security, particularly with such sensitive data being used, and minimizing potential to propagate biases in AI systems [48,49]. Language features were excluded where all values were zero. For models classifying interaction themes, original CICS codes were converted to *problem or context description* interactions versus other interaction themes combined.

Coproduced Linguistic Feature Extraction

The autoCICS development team was deliberately assembled to ensure that it comprised key research and clinical stakeholders with regard to the characteristics of an automated psychological therapy rating tool. The team comprised 2 psychological therapists and a psychological therapy assistant (SM, NM, and CM, respectively), who offered clinical expertise; 3 service-user researchers (MR, FH, and DW), who offered patient-related knowledge and experience; an applied linguist (DH), who contributed expertise on linguistic functions and patterns; an AI ethics researcher (EPV); and an explainable AI researcher (JC), who added an understanding of how participatory methodology could be meaningfully translated into NLP features. The team members were separately surveyed about what language markers in patient-therapist interactions they thought might be indicative of greater patient activation—that is, active engagement, involvement, and ownership of the therapeutic process. The team members were also asked what language markers they felt might indicate a patient's disengagement and withdrawal from therapeutic processes. The features identified were then collaboratively translated into NLP features by three other team members: an NLP researcher (YL) and two digital research experts (TJ and GF). Table 3 presents examples of the language features suggested by different disciplinary groups within the team (refer to Multimedia Appendix 1 for the final language features used in validation with nonsignificant features removed). This process aimed to generate understandable language features from different relevant perspectives for the future product's end users. This methodology aimed to enhance transparency and involve domain experts in selecting input features rather than unsupervised learning from the data, which would likely be less interpretable. Language features were extracted using the Python Natural Language Toolkit (NLTK Project) and the Python library, TextBlob.

Table 3. Examples of suggested language features deemed indicative of greater patient activation.

Suggestion source and language feature	Related study
Service users	
Less profanity (swear words and curses)	Coppersmith et al [50]
Fewer absolutes (always, never, and everything)	Al Mosaiwi and Johnstone [51]
Fewer maximizers (worst and most)	Strohm and Klinger [52]
Psychological therapists	
Positive sentiment (happy, glad, and good)	Calvo et al [53]
Intensity of positive sentiment (polarity and frequency)	Calvo et al [53]
Lower ratio of illness: wellness terminology	Arseniev et al [54]
Applied linguist	
Fewer deontics (eg, must, should, and ought)	Van der Zanden [55]
Fewer qualifier words (eg, but and though)	Jeong [56]
Ratio of plural: singular first-person pronouns	Rude et al [57]
Explainable artificial intelligence researcher	
Longer sentences (number of words)	Hirschberg et al [58]
Longer words (number of characters)	Pestian et al [59]
Lower Flesch-Kincaid readability score (more complex sentences)	Pestian et al [59]

Machine Learning

A bagged trees algorithm was used to classify patient activation level, that is, differentiating between interactions rated positively (+1 or +2) and those rated negatively or neutral (−1, −2, or 0). The model used a constant weight of 3 for misclassified instances at level 1 to penalize misclassifications in the less frequent class. The constant of 3 was reached through algorithm optimization during training. A k-nearest neighbor algorithm was used to classify interaction types; specifically, differentiating between *problem and context description* interactions and other interaction types, given the prognostic importance of these interactions. Both models were developed using MATLAB (version 2021a; MathWorks, Inc). The standard implementation from MATLAB uses hyperparameter tuning intrinsically. Exploratory modeling also evaluated the classification of other, less frequent interaction types rated on the CICS (eg, *evaluations of self or therapy*). The synthetic minority oversampling technique [60] was initially applied to augment the data, but it did not significantly improve the results;

therefore, it was removed, particularly given that highly unbalanced data set and potential clinical use.

Results

Sample Characteristics

The included participants were predominantly White British (40/53, 75%), and three-quarters (40/53, 75%) were female. All participants had been assessed as experiencing severe health anxiety using the short health anxiety inventory, but all participants reported additional comorbidities. On average, participants met criteria for 7 (SD 3.7) mental disorders from the structured clinical interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, assessment, most commonly generalized anxiety disorder. Participants also reported a mean 1 (SD 1.15) additional chronic physical health condition, most commonly chronic pain (refer to Table 4 for participant demographics and clinical characteristics).

Table 4. Demographics and clinical characteristics of participants (N=53).

Variable	Values
Demographics	
Sex, female, n (%)	40 (75)
Age (years), mean (SD)	36 (15)
Ethnicity, n (%)	
White British	40 (75)
Other	13 (24)
Unemployed, n (%)	6 (11)
Clinical characteristics	
SCID[a] diagnoses, mean (SD; range)	7 (3.7; 0-16)
Generalized anxiety disorder, n (%)	35 (66)
Hypochondriasis, n (%)	34 (64)
Somatoform disorders, n (%)	33 (62)
Current depressive episode, n (%)	32 (60)
Panic disorder, n (%)	32 (60)
Long-term physical health problems, mean (SD; range)	1 (1.15; 0-6)
Chronic pain, n (%)	13 (25)
Chronic fatigue, n (%)	5 (9)
Functional neurological disorders, n (%)	5 (9)
Irritable bowel syndrome, n (%)	4 (8)
Arthritis, n (%)	4 (8)
Diabetes, n (%)	4 (8)

[a] SCID: structured clinical interview for the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.

Data Characteristics

Problem or context description interactions were the most commonly coded CICS theme, accounting for 54.6% (22,967/42,064) of the interactions in the training data set and 46.8% (4324/9239) of the interactions in the test data set. Conversely, interactions involving patients' *evaluations of self or therapy* were the least coded interaction type, accounting for 2.4% (1010/42,064) and 3% (277/9239) of the training data set and test data set, respectively.

Interaction Classification

Given that the data set was imbalanced, *F*-scores are reported alongside accuracy scores because they are less sensitive to class imbalance. For the model based on a k-nearest neighbor algorithm used to identify CICS-rated interaction themes (correctly identifying *problem or context description* interactions versus other interactions), an overall accuracy of 73% (precision=82%, recall=80%, and *F*-score=73%) was observed in the test data set. The model used to classify the CICS-rated patient activation level (positive versus negative or neutral) obtained an 81% accuracy (precision=87%, recall=75%, and *F*-score=87%) in the test data set.

Exploratory models aiming to classify less frequent interaction themes (*action planning and idea generation*, *evaluations of self or therapy*, *information discussion*, *noticing change or otherwise*, *problem analysis or understanding*, and *structuring and task focus*) obtained lower-than-average *F*-scores of 20% because of very high class imbalance.

Discussion

Principal Findings

This study indicates that collaboratively and transparently developed AI can be used to discriminate between high and low patient activation from turns of speech in psychological therapy sessions. The language features used also discriminated between *problem or context description interactions* and other interaction types. However, the model could not discriminate among other interaction types on the CICS (eg, *action planning* versus *problem analysis or understanding*). The codevelopment approach applied may help to improve trust in the decisions made by an autoCICS psychological therapy rating tool among end users, including patients, psychological therapists, and service managers [31]. The model was also enhanced by including key stakeholders in the selection of language features that formed the basis of the prediction models, rather than using an exclusively data-driven approach likely to end in more opaque and potentially spurious processes that have reduced trust in AI generally [48]. The involvement of stakeholders in this way also helps to develop a fit-for-purpose system within health care when AI applications often lack adequate end-user involvement [61]. Overall, the findings suggest that reasonable predictive accuracy was achieved with the participatory methodology applied (involving key stakeholders in the AI model development).

Comparison With Prior Work

By including participatory approaches to enhance trust and interpretability, this study builds on existing research where AI has been used to automate psychological therapy rating tools [10,62]. Similar levels of agreement with human rating reliability were achieved in this study compared with previous attempts to automate psychological therapy turn-by-turn ratings [9,63]. This suggests that the simplifications made to the modeling for greater interpretability have not been excessively detrimental to model performance. An automated assessment that takes account of both therapist and patient utterances in this study may also help build a clearer understanding of language features associated with therapeutic responsiveness in future [47]. This is particularly relevant because many current machine learning models focus on either therapist or patient utterances alone [9,11]. Whereas most previous supervised learning models have focused on in-session behaviors related to a specific therapeutic model (eg, motivational interviewing [10]), the autoCICS in this study assesses patient activation—a construct that may have relevance across psychological therapy models and treatments in other domains [64]. Furthermore, this study expands the range of patients included in this type of modeling with a patient sample experiencing multimorbidity at baseline. Given the importance of health care professionals' interaction style and responsiveness to enhance patient activation during consultations with people experiencing multimorbidity, an automated interaction assessment has potential for broad application in improving care [21]. By including the now often used modality of remote psychological therapy, this study also expands the range of psychological therapy delivery modalities where NLP has been applied.

Limitations

This study used a relatively small sample size for machine learning studies. This means that the breadth of interaction types and language features used may be restricted, making the results less generalizable. However, the sample size is typical compared with previous studies of NLP in psychological therapy [65]. The smaller sample size also limited use of more complex modeling methods that could have improved classification precision and sensitivity, especially when considering more levels of granularity with regard to the interaction types and patient activation levels. Relatedly, a limited number of therapists were included in the data set; a more representative sample of therapists may have helped identify and define important differences among therapists who could be included in models to improve accuracy. A larger number of therapists could also help to discriminate among different clustered therapist phenotypes, where different interaction styles could be attributed to specific therapist groups.

In exploratory modeling, the classifier accuracy in less frequent classes of interaction was much lower. This suggests that either there was insufficient data to train the model, or the language features applied in the models did not discriminate among these interaction themes very well. The result is that the current classifier could not offer refined, granular feedback to practitioners on more detailed aspects of their session contents. Another possible explanation for the classifier's poor performance in discriminating among different interaction types (eg, structuring interactions versus information giving) is that the same language features were used to classify both patient activation level and interaction type. Different language features may have given clearer differentiation on interaction types.

Although the CICS-labeled data used to train the model in this study aimed to address therapist responsiveness by combined ratings of therapist and patient data, this prevents an understanding of individual contributions to patient activation from either therapist or patient; for example, where a patient's interaction indicates movement toward greater engagement, but the therapist's response undermines this. The current classification process would struggle to identify these occasions, which could be important for therapist feedback.

Although this study indicates that the autoCICS achieved good discriminative validity, it is unclear whether this would be sufficiently accurate for reliable use in clinical settings. Furthermore, the practical, clinical value of the classifier would need to be evaluated in practice before significance could be assessed. Therefore, further model validation is required, and the feasibility and acceptability of the tool in clinical practice should be assessed, given the catalog of implementation failures for AI tools in health care more broadly [24].

Future Research

The automated ratings presented in this paper require external validation to clarify whether interactions rated as high in patient activation associate with assessments of patient activation used in clinical practice, such as the PAM, conducted at the same time point. The clinical utility of the automated assessment cannot be assured until such validation has been carried out.

Larger-scale validation could use a varied, more representative patient and therapist sample to help improve the generalizability of the model and address potential biases in model decisions. Future research may also benefit from use of routine care data sets (in contrast to research trial data, as in this study). This may give a closer representation of therapeutic processes experienced in real-world therapy and, therefore, increase wider applicability. Validation across different psychological therapy models and presenting problems would also help to establish transferable aspects of the model's utility. Future research should also clarify the prognostic value of the autoCICS not only to establish whether sufficient reliability has been achieved to retain the CICS predictive validity but also to assess whether predictive validity can be improved using a codevelopment approach.

This study, alongside most previous research, has focused on lexical elements of psychological therapy content (transcribed words), but it does not address the nonlexical, phonological features of talk (such as intonation and prosody) that can be an important predictor of health [66]. Therefore, future research should address the integration of lexical and phonological analyses of psychological therapy content for more accurate representations of in-session events. Finally, future research should identify means of building and maintaining codevelopment, interpretability, and transparency within more complex AI analyses of psychological therapy content. Collaboratively developed models may not identify the same features as either expert-designed models or unsupervised learning models, but they may be more trustworthy and fit for purpose for end users [29]. In future, contrasting results from participatory approaches, such as the one used in this study, with more *black box* approaches to developing an automated classifier would give an informed view on the trade-off between model accuracy and algorithmic trust. This will be particularly important if greater accuracy is to be achieved in classifying more detailed interaction types, which could not be achieved with the current methodology. Importantly, the participatory methods used do not preclude the use of more complex algorithms to develop models in future research.

Clinical Implications

This study presents the initial development of an automated assessment of patient activation that can be rated turn by turn routinely in psychological therapy. Alongside other advances, this methodology may help enhance deliberate practice techniques in psychological therapy. Deliberate practice aims to identify therapeutic microskills requiring improvement and then improve these skills through corrective practice [16]. In conjunction with a further developed autoCICS, alongside associated training and supervision, therapists could learn to recognize problematic patterns more easily and practice different responses.

Conclusions

A participatory methodology was applied to develop a novel approach for the assessment of in-session patient activation during psychological therapy. This approach can support the responsible design and development of autonomous and intelligent systems in psychological therapy by building trust among stakeholders from initial development.

Language features identified by a multiperspective stakeholder collaboration can be used to discriminate between high and low patient activation in psychological therapy session contents but were limited in their ability to discriminate among different psychological therapy interaction types. However, larger-scale replication is required before clinical utility can be assessed.

Acknowledgments

This work is funded by the Engineering and Physical Sciences Research Council's Human Data Interaction Network "Future of Mental Health," with support from Health Data Research UK (University of Nottingham). This study was supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration East Midlands Development and Skills Enhancement Award (SM, NIHR300822) and Digital Research Service, University of Nottingham. EPV also acknowledges the financial support of the NIHR Nottingham Biomedical Research Centre and UK Research and Innovation. UK Research and Innovation does not necessarily endorse the views expressed by the author. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Final language features used in modeling.
[DOCX File , 14 KB - medinform_v10i11e38168_app1.docx]

References

1. Barkham M, Lutz W, Lambert M, Saxon D. Therapist effects, effective therapists, and the law of variability. In: How and Why Are Some Therapists Better Than Others?: Understanding Therapist Effects. Washington, D.C., United States: American Psychological Association; 2017.
2. Baldwin S, Imel Z. Therapist effects: findings and methods. In: Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. New Jersey: Wiley; 2013.

3. Goldberg SB, Rousmaniere T, Miller SD, Whipple J, Nielsen SL, Hoyt WT, et al. Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. J Couns Psychol 2016 Jan;63(1):1-11. [doi: 10.1037/cou0000131] [Medline: 26751152]
4. Johnsen TJ, Friborg O. The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: a meta-analysis. Psychol Bull 2015 Jul;141(4):747-768. [doi: 10.1037/bul0000015] [Medline: 25961373]
5. Prochaska JO, Norcross JC, Saul SF. Generating psychotherapy breakthroughs: transtheoretical strategies from population health psychology. Am Psychol 2020 Oct;75(7):996-1010. [doi: 10.1037/amp0000568] [Medline: 31763861]
6. Perepletchikova F. On the topic of treatment integrity. Clin Psychol (New York) 2011 Jun;18(2):148-153 [FREE Full text] [doi: 10.1111/j.1468-2850.2011.01246.x] [Medline: 21769167]
7. Waller G, Turner H. Therapist drift redux: why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. Behav Res Ther 2016 Feb;77:129-137. [doi: 10.1016/j.brat.2015.12.005] [Medline: 26752326]
8. Samoili S, López CM, Gómez E, De PG, Martínez-Plumed F, Delipetrev BA. AI Watch. Defining Artificial Intelligence. Towards An Operational Definition and Taxonomy of Artificial Intelligence. Luxembourg: Publications Office of the European Union; 2020.
9. Ewbank MP, Cummins R, Tablan V, Catarino A, Buchholz S, Blackwell AD. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: a deep learning approach to automatic coding of session transcripts. Psychother Res 2021 Mar 03;31(3):326-338. [doi: 10.1080/10503307.2020.1788740] [Medline: 32619163]
10. Atkins DC, Steyvers M, Imel ZE, Smyth P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implement Sci 2014 Apr 24;9(1):49 [FREE Full text] [doi: 10.1186/1748-5908-9-49] [Medline: 24758152]
11. Ewbank MP, Cummins R, Tablan V, Bateup S, Catarino A, Martin AJ, et al. Quantifying the association between psychotherapy content and clinical outcomes using deep learning. JAMA Psychiatry 2020 Jan 01;77(1):35-43 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.2664] [Medline: 31436785]
12. Chikersal P, Belgrave D, Doherty G, Enrique A, Palacios J, Richards D, et al. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20: CHI Conference on Human Factors in Computing Systems; Apr 25 - 30, 2020; Honolulu HI USA. [doi: 10.1145/3313831.3376341]
13. Moyers T, Martin T, Catley D, Harris KJ, Ahluwalia JS. Assessing the integrity of motivational interviewing interventions: reliability of the motivational interviewing skills code. Behav Cognit Psychother 2003 May;31(2):177-184. [doi: 10.1017/s1352465803002054]
14. McKay JR. Lessons learned from psychotherapy research. Alcohol Clin Exp Res 2007 Oct;31(10 Suppl):48s-54s. [doi: 10.1111/j.1530-0277.2007.00493.x] [Medline: 17880346]
15. Glazebrook C, Davies EB. Outcome feedback technology helps therapists to tailor care. Lancet Psychiatry 2018 Jul;5(7):529-531. [doi: 10.1016/s2215-0366(18)30212-8]
16. Miller S, Hubble M, Chow D. Better Results Using Deliberate Practice to Improve Therapeutic Effectiveness. Washington, DC: American Psychological Association; 2020.
17. Rousmaniere T, Goodyear R, Miller S, Wampold B. Improving psychotherapy outcomes guidelines for making psychotherapist expertise development routine and expected. In: The Cycle of Excellence: Using Deliberate Practice to Improve Supervision and Training. Chichester, UK: Wiley; 2017.
18. Johns RG, Barkham M, Kellett S, Saxon D. A systematic review of therapist effects: a critical narrative update and refinement to review. Clin Psychol Rev 2019 Feb;67:78-93. [doi: 10.1016/j.cpr.2018.08.004] [Medline: 30442478]
19. Stafford M, Steventon M, Thorlby R, Fisher R, Turton C, Deeny S. Briefing: understanding the health care needs of people with multiple health conditions. The Health Foundation. 2018 Nov. URL: https://tinyurl.com/2x3xw52u [accessed 2022-02-11]
20. Reducing emergency admissions: unlocking the potential of people to better manage their long-term conditions. National Grey Literature Collection. 2018. URL: https://www.health.org.uk/sites/default/files/Reducing-Emergency-Admissions-long-term-conditions-briefing.pdf [accessed 2022-03-08]
21. Mercer SW, Fitzpatrick B, Guthrie B, Fenwick E, Grieve E, Lawson K, et al. The CARE Plus study - a whole-system intervention to improve quality of life of primary care patients with multimorbidity in areas of high socioeconomic deprivation: exploratory cluster randomised controlled trial and cost-utility analysis. BMC Med 2016 Jun 22;14(1):88 [FREE Full text] [doi: 10.1186/s12916-016-0634-2] [Medline: 27328975]
22. Derksen F, Bensing J, Lagro-Janssen A. Effectiveness of empathy in general practice: a systematic review. Br J Gen Pract 2013 Jan 01;63(606):e76-e84. [doi: 10.3399/bjgp13x660814]
23. Wind T, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the 'black swan' for mental health care and a turning point for e-health. Internet Interv 2020 Apr;20:100317 [FREE Full text] [doi: 10.1016/j.invent.2020.100317] [Medline: 32289019]
24. Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. NPJ Digit Med 2019 Aug 16;2(1):77 [FREE Full text] [doi: 10.1038/s41746-019-0155-4] [Medline: 31453372]

25. Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 2017 Aug 15;31(4):611-627. [doi: 10.1007/s13347-017-0279-x]
26. Carr S. 'AI gone mental': engagement and ethics in data-driven technology for mental health. J Ment Health 2020 Apr 30;29(2):125-130. [doi: 10.1080/09638237.2020.1714011] [Medline: 32000544]
27. Balaram B, Greenham T, Leonard J. Artificial Intelligence: real public engagement. RSA. 2018 May 30. URL: https://tinyurl.com/4ajzu5e5 [accessed 2022-06-07]
28. Owen R, von Schomberg R, Macnaghten P. An unfinished journey? Reflections on a decade of responsible research and innovation. J Responsible Innov 2021 Jul 26;8(2):217-233. [doi: 10.1080/23299460.2021.1948789]
29. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]
30. Ntoutsi E. Bias in AI-systems: a multi-step approach. In: Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence. 2020 Presented at: 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence; Nov, 2020; Dublin, Ireland.
31. Ehsan U, Liao Q, Muller M, Riedl M, Weisz J. Expanding explainability: towards social transparency in AI systems. arXiv 2021 [FREE Full text] [doi: 10.1145/3411764.3445188]
32. Malins S, Moghaddam N, Morriss R, Schröder T, Brown P, Boycott N, et al. Patient activation in psychotherapy interactions: developing and validating the consultation interactions coding scheme. J Clin Psychol 2020 Apr;76(4):646-658 [FREE Full text] [doi: 10.1002/jclp.22910] [Medline: 31825098]
33. Hibbard JH, Greene J. What the evidence shows about patient activation: better health outcomes and care experiences; fewer data on costs. Health Aff (Millwood) 2013 Feb;32(2):207-214. [doi: 10.1377/hlthaff.2012.1061] [Medline: 23381511]
34. Mosen DM, Schmittdiel J, Hibbard J, Sobel D, Remmers C, Bellows J. Is patient activation associated with outcomes of care for adults with chronic conditions? J Ambul Care Manage 2007;30(1):21-29. [doi: 10.1097/00004479-200701000-00005] [Medline: 17170635]
35. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res 2005 Dec;40(6 Pt 1):1918-1930 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00438.x] [Medline: 16336556]
36. Graffigna G, Barello S, Bonanomi A, Lozza E. Measuring patient engagement: development and psychometric properties of the Patient Health Engagement (PHE) Scale. Front Psychol 2015 Mar 27;6:274 [FREE Full text] [doi: 10.3389/fpsyg.2015.00274] [Medline: 25870566]
37. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004 Aug;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
38. Deen D, Lu W, Rothstein D, Santana L, Gold MR. Asking questions: the effect of a brief intervention in community health centers on patient activation. Patient Educ Couns 2011 Aug;84(2):257-260. [doi: 10.1016/j.pec.2010.07.026] [Medline: 20800414]
39. Armstrong N, Tarrant C, Martin G, Manktelow B, Brewster L, Chew S. Independent evaluation of the feasibility of using the Patient Activation Measure in the NHS in England. NHS England. 2016 Apr 25. URL: https://www.england.nhs.uk/wp-content/uploads/2016/04/pa-interim-report-summary.pdf [accessed 2022-06-07]
40. Malins S, Moghaddam N, Morriss R, Schröder T, Brown P, Boycott N. Predicting outcomes and sudden gains from initial in-session interactions during remote cognitive-behavioural therapy for severe health anxiety. Clin Psychol Psychother 2021 Jul 06;28(4):891-906. [doi: 10.1002/cpp.2543] [Medline: 33368731]
41. Malins S, Moghaddam N, Morriss R, Schröder T, Brown P, Boycott N. The predictive value of patient, therapist, and in-session ratings of motivational factors early in remote cognitive behavioural therapy for severe health anxiety. Br J Clin Psychol 2022 Jun 12;61(2):364-384. [doi: 10.1111/bjc.12328] [Medline: 34514604]
42. Morriss R, Patel S, Malins S, Guo B, Higton F, James M, et al. Clinical and economic outcomes of remotely delivered cognitive behaviour therapy versus treatment as usual for repeat unscheduled care users with severe health anxiety: a multicentre randomised controlled trial. BMC Med 2019 Jan 23;17(1):16 [FREE Full text] [doi: 10.1186/s12916-019-1253-5] [Medline: 30670044]
43. First MB. Structured Clinical Interview for DSM-IV Axis I Disorders : Patient Edition (February 1996 Final), SCID-I/P. New York: New York State Psychiatric Institute; 1998.
44. Greene J, Hibbard JH, Sacks R, Overton V, Parrotta CD. When patient activation levels change, health outcomes and costs change, too. Health Aff (Millwood) 2015 Mar;34(3):431-437. [doi: 10.1377/hlthaff.2014.0452] [Medline: 25732493]
45. Malins S, Moghaddam N, Morriss R, Schroder T, Cope N, Brown P. Consultation Interaction Coding Scheme (CICS) 1.6. figshare. Figshare. 2018. URL: https://figshare.com/articles/Consultation_Interaction_Coding_Scheme_CICS_/7302386 [accessed 2022-06-07]
46. Kramer U, Stiles WB. The responsiveness problem in psychotherapy: a review of proposed solutions. Clin Psychol Sci Pract 2015 Sep;22(3):277-295. [doi: 10.1111/cpsp.12107]
47. Stiles W, Horvath A. Appropriate responsiveness as a contribution to therapist effects. In: How and Why Are Some Therapists Better Than Others?: Understanding Therapist Effects. Washington, DC: American Psychological Association; 2017.

48. What consumers really think about AI: a global study. Pegasystems. 2018. URL: https://www.ciosummits.com/what-consumers-really-think-about-ai.pdf [accessed 2022-06-07]
49. Liu B, Ding M, Zhu T, Xiang Y, Zhou W. Adversaries or allies? Privacy and deep learning in big data era. Concurrency Computat Pract Exper 2018 Dec 14;31(19):e5102. [doi: 10.1002/cpe.5102]
50. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014 Presented at: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; Jun, 2014; Baltimore, Maryland, USA. [doi: 10.3115/v1/w14-3207]
51. Al-Mosaiwi M, Johnstone T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin Psychol Sci 2018 Jul 05;6(4):529-542 [FREE Full text] [doi: 10.1177/2167702617747074] [Medline: 30886766]
52. Strohm F, Klinger R. An empirical analysis of the role of amplifiers, downtoners, and negations in emotion classification in microblogs. In: Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). 2018 Presented at: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); Oct 01-03, 2018; Turin, Italy. [doi: 10.1109/dsaa.2018.00087]
53. CALVO RA, MILNE DN, HUSSAIN MS, CHRISTENSEN H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng 2017 Jan 30;23(5):649-685. [doi: 10.1017/s1351324916000383]
54. Arseniev-Koehler A, Mozgai S, Scherer S. What type of happiness are you looking for? - A closer look at detecting mental health from language. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. 2018 Presented at: Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; Jun, 2-18; New Orleans, LA. [doi: 10.18653/v1/w18-0601]
55. Van der Zanden R, Curie K, Van Londen M, Kramer J, Steen G, Cuijpers P. Web-based depression treatment: associations of clients' word use with adherence and outcome. J Affect Disord 2014 May;160:10-13 [FREE Full text] [doi: 10.1016/j.jad.2014.01.005] [Medline: 24709016]
56. Jeong AC. The effects of linguistic qualifiers and intensifiers on group interaction and performance in computer-supported collaborative argumentation. Int Rev Res Open Distributed Learn 2006 Feb 22;6(3). [doi: 10.19173/irrodl.v6i3.258]
57. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cognit Emotion 2004 Dec;18(8):1121-1133. [doi: 10.1080/02699930441000030]
58. Hirschberg J, Hjalmarsson A, Elhadad N. “you’re as sick as you sound”: using computational approaches for modeling speaker state to gauge illness and recovery. In: Advances in Speech Recognition. Boston, MA: Springer; 2010.
59. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: a content analysis. Biomed Inform Insights 2010 Aug 04;2010(3):19-28 [FREE Full text] [doi: 10.4137/bii.s4706] [Medline: 21643548]
60. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artificial Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
61. Fischer G. End-user development: empowering stakeholders with artificial intelligence, meta-design, and cultures of participation. In: End-User Development. Cham: Springer; 2021.
62. Goldberg SB, Tanana M, Imel ZE, Atkins DC, Hill CE, Anderson T. Can a computer detect interpersonal skills? Using machine learning to scale up the Facilitative Interpersonal Skills task. Psychother Res 2021 Mar 16;31(3):281-288 [FREE Full text] [doi: 10.1080/10503307.2020.1741047] [Medline: 32172682]
63. Tanana M, Hallgren KA, Imel ZE, Atkins DC, Srikumar V. A comparison of natural language processing methods for automated coding of motivational interviewing. J Subst Abuse Treat 2016 Jun;65:43-50 [FREE Full text] [doi: 10.1016/j.jsat.2016.01.006] [Medline: 26944234]
64. Hibbard J, Gilburt H. Supporting People to Manage Their Health An Introduction to Patient Activation. London, UK: The King's Fund; 2014.
65. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. Psychother Res 2021 Jan 29;31(1):92-116. [doi: 10.1080/10503307.2020.1808729] [Medline: 32862761]
66. Sertolli B, Ren Z, Schuller BW, Cummins N. Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. Comput Speech Language 2021 Jul;68:101204. [doi: 10.1016/j.csl.2021.101204]

Abbreviations

AI: artificial intelligence
autoCICS: automated consultation interactions coding scheme
CICS: consultation interactions coding scheme
NLP: natural language processing
PAM: patient activation measure
rCBT: remotely delivered cognitive behavioral therapy

Edited by T Hao; submitted 21.03.22; peer-reviewed by J Hibbard, H Burkhardt, A Teles; comments to author 26.04.22; revised version received 07.06.22; accepted 27.06.22; published 08.11.22.

Please cite as:
Malins S, Figueredo G, Jilani T, Long Y, Andrews J, Rawsthorne M, Manolescu C, Clos J, Higton F, Waldram D, Hunt D, Perez Vallejos E, Moghaddam N
Developing an Automated Assessment of In-session Patient Activation for Psychological Therapy: Codevelopment Approach
JMIR Med Inform 2022;10(11):e38168
URL: https://medinform.jmir.org/2022/11/e38168
doi:10.2196/38168
PMID:36346654

Original Paper

A Transfer Learning Approach to Correct the Temporal Performance Drift of Clinical Prediction Models: Retrospective Cohort Study

Xiangzhou Zhang[1*], PhD; Yunfei Xue[1,2*], MSc; Xinyu Su[1,2*], MSc; Shaoyong Chen[1,2], MSc; Kang Liu[1,3], PhD; Weiqi Chen[1], PhD; Mei Liu[4], PhD; Yong Hu[1], PhD

[1]Big Data Decision Institute, Jinan University, Guangzhou, China

[2]College of Information Science and Technology, Jinan University, Guangzhou, China

[3]School of Management, Jinan University, Guangzhou, China

[4]Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States

[*]these authors contributed equally

Corresponding Author:
Yong Hu, PhD
Big Data Decision Institute
Jinan University
601 Huangpu Road West
Guangzhou, 510632
China
Phone: 86 02085223261
Email: henryhu200211@163.com

Abstract

Background: Clinical prediction models suffer from performance drift as the patient population shifts over time. There is a great need for model updating approaches or modeling frameworks that can effectively use the old and new data.

Objective: Based on the paradigm of transfer learning, we aimed to develop a novel modeling framework that transfers old knowledge to the new environment for prediction tasks, and contributes to performance drift correction.

Methods: The proposed predictive modeling framework maintains a logistic regression–based stacking ensemble of 2 gradient boosting machine (GBM) models representing old and new knowledge learned from old and new data, respectively (referred to as transfer learning gradient boosting machine [TransferGBM]). The ensemble learning procedure can dynamically balance the old and new knowledge. Using 2010-2017 electronic health record data on a retrospective cohort of 141,696 patients, we validated TransferGBM for hospital-acquired acute kidney injury prediction.

Results: The baseline models (ie, transported models) that were trained on 2010 and 2011 data showed significant performance drift in the temporal validation with 2012-2017 data. Refitting these models using updated samples resulted in performance gains in nearly all cases. The proposed TransferGBM model succeeded in achieving uniformly better performance than the refitted models.

Conclusions: Under the scenario of population shift, incorporating new knowledge while preserving old knowledge is essential for maintaining stable performance. Transfer learning combined with stacking ensemble learning can help achieve a balance of old and new knowledge in a flexible and adaptive way, even in the case of insufficient new data.

(JMIR Med Inform 2022;10(11):e38053) doi:10.2196/38053

KEYWORDS

transfer learning; clinical prediction model; performance drift; concept drift; acute kidney injury

Introduction

Clinical risk prediction models can provide decision-making support on therapeutic interventions and resource allocation, and thus can improve patient outcomes and reduce medical costs [1]. Along with the increasing availability and volume of electronic health record (EHR) data, these models are evolving from rule-based to data-driven probability-based tools, for example, machine learning–based patient outcome prediction models [2]. One of the critical challenges is performance drift over time, which results from either gradual or quick data shifts in the patient population, such as changing patient outcome rate, evolving clinical practices, and improving measurement accuracy [3].

To correct temporal performance drift, a range of model updating approaches are available, including recalibration, model-specific adaptation (eg, reweighting the leaf nodes of each tree in a random forest [RF] model and an incremental learning method for a neural network model), model extension (eg, incorporating new predictors), and full model refitting [1]. These updating approaches vary in analytical complexity, old data and updated sample requirements, and computational demands. Usually, full model refitting is not the leading choice, especially in clinical use, owing to the risk of overfitting when new (and often smaller) data are used alone, while old data are completely discarded [1]. The essence of model updating is to create models that are constantly updated and adapted to the new incoming data, while balancing between both new and old knowledge [4-7].

Acute kidney injury (AKI) is a potentially life-threatening clinical syndrome, for which the only effective treatments are supportive care and dialysis, and it affects 10%-15% of all inpatients and more than 50% of critical care patients, and results in high mortality [8,9]. For AKI prediction, Davis et al [2] developed 7 common regression and machine learning models, and found that discrimination performance declines were statistically significant but small for all models. Since they collected data solely from US Department of Veterans Affairs hospitals, it is not a typical scenario of population drift. Using data collected from Royal London Hospital, which hosts Europe's largest kidney treatment facility, Haines et al [10] developed risk prediction models for AKI after trauma, with the area under the receiver operating characteristic curve (AUROC) declining from 0.77 (0.72-0.81) in the development set (February 2012 to October 2014) to 0.70 (0.64-0.77) in the validation set (November 2014 to May 2016), and significant temporal performance drift.

In this study, we developed a clinical risk prediction model for hospital-acquired AKI. The model has been named transfer learning gradient boosting machine (TransferGBM), which is based on a transfer learning paradigm and maintains a stacking ensemble of 2 base gradient boosting machine (GBM) learners. Transfer learning has been proven to be one of the most effective ways to deal with data scarcity (eg, in the scenario where new data are not sufficient or available at a low cost) and data distribution discrepancies in many areas [11-17]. Transfer learning aims to selectively reuse data or knowledge from the source domain to assist the modeling process on the target domain, and it can be used to tackle the performance drift problem by regarding the old data as the source domain and the new data as the target domain. Since existing transfer learning approaches focus on optimizing performance only in the target domain, we still need a well-designed mechanism to incorporate and balance the old and new knowledge learned from the source and target domains.

Methods

Definition of AKI

According to the Kidney Disease Improving Global Outcomes (KDIGO) clinical practice guidelines for AKI, we adopted serum creatinine (SCr)-based criteria to stage the severity of AKI [18]. We did not use urine output to define AKI because it is less likely to be accurate outside the critical care environment [19,20]. Mild AKI ("AKI stage 1") is defined as an increase in SCr of 1.5 to 1.9 times the baseline value within 7 days or an increase in SCr to 0.3 mg/dL (26.5 µmol/L) or more within 48 hours. The baseline creatinine value is defined as the most recent SCr if available; otherwise, it is the admission SCr. Moderate AKI ("AKI stage 2") is defined as an increase in SCr of 2.0 to 2.9 times the baseline value within 7 days. The most severe AKI ("AKI stage 3") is defined as an increase in SCr of 3.0 or more times the baseline value within 7 days or an increase in SCr to 4 mg/dL (353.5 µmol/L) after an acute increase of at least 0.3 mg/dL within 48 h or initiation of renal replacement therapy.

Study Cohort

The study constructed a retrospective cohort using deidentified EHR data from 2010 to 2017 in the University of Kansas Medical Center. The data have been used in a previous study [20] including a total of 141,696 adult patients (121,537 non-AKI patients; 20,159 any AKI patients; 3150 AKI stage ≥2 patients; and 1491 AKI stage 3 patients). To reflect the inpatient population shift, patients enrolled in different years were regarded as distinct individuals (ie, we handled the data at the patient-encounter level).

As shown in Table 1, the proportion of elderly patients (ie, age ≥65) generally increased every year, from 31.7% in 2010 to 36.5% in 2017. The proportion of patients between the ages of 46 and 55 years decreased every year, while the proportion of patients in other age groups remained the same. The ratio of male to female patients did not change much over time, and was basically maintained at 1:1. The proportion of White patients always ranked first, accounting for more than 70% of the total number of samples in each year, while the proportion of Native Hawaiians was the least (only 0.1%). Only the proportion of patients from different ethnicities remained stable over time, without obvious changes. The proportion of African Americans was more in 2010 than in all other years, and the proportion of White patients was slightly less in 2010 than in all other years. In addition, the incidence of AKI (any AKI) showed a clear downward trend, from 16.9% in 2010 to 12.8% in 2017.

Table 1. Demographic information.

Feature	Year							
	2010 (N=14,946)	2011 (N=15,422)	2012 (N=16,682)	2013 (N=17,450)	2014 (N=18,701)	2015 (N=20,094)	2016 (N=20,399)	2017 (N=18,002)
Age group (years), n (%)								
18-25	869 (5.8)	886 (5.7)	923 (5.5)	918 (5.3)	1077 (5.8)	1082 (5.4)	1086 (5.3)	1001 (5.6)
26-35	1290 (8.6)	1275 (8.3)	1468 (8.7)	1567 (9.0)	1717 (9.7)	1814 (9.0)	1823 (8.9)	1664 (9.2)
36-45	1640 (11.0)	1727 (11.2)	1696 (10.2)	1861 (10.7)	1819 (9.7)	2136 (10.6)	2196 (10.8)	1919 (10.7)
46-55	3025 (20.2)	2998 (19.4)	3203 (19.2)	3133 (19.0)	3150 (16.8)	3482 (17.3)	3259 (16.0)	2762 (15.3)
56-65	3383 (22.6)	3659 (23.7)	3951 (23.7)	4161 (23.8)	4558 (24.4)	4897 (24.4)	4840 (23.7)	4088 (22.7)
>65	4739 (31.7)	4877 (31.6)	5441 (32.6)	5810 (33.3)	6380 (34.1)	6683 (33.3)	7195 (35.3)	6568 (36.5)
Sex, n (%)								
Male	7547 (50.5)	7635 (49.5)	8432 (50.5)	8640 (49.5)	9307 (49.8)	10,114 (50.3)	10,250 (50.2)	9045 (50.2)
Female	7399 (49.5)	7787 (50.5)	8250 (49.5)	8810 (50.5)	9394 (50.2)	9980 (49.7)	10,149 (49.8)	8957 (49.8)
Race, n (%)								
American Indian	53 (0.4)	52 (0.3)	46 (0.3)	79 (0.5)	68 (0.4)	87 (0.4)	80 (0.4)	63 (0.3)
Asian	125 (0.8)	128 (0.8)	153 (0.9)	167 (1.0)	210 (1.1)	184 (0.9)	254 (1.2)	149 (0.8)
African American	2286 (15.3)	2240 (14.5)	2255 (13.5)	2510 (13.4)	2685 (14.4)	2883 (14.3)	2896 (14.2)	2614 (14.5)
Native Hawaiian	11 (0.1)	20 (0.1)	9 (0.1)	9 (0.1)	15 (0.1)	10 (0.1)	18 (0.1)	14 (0.1)
White	10,915 (72.9)	11,485 (74.5)	12,691 (76.1)	13,331 (76.4)	14,322 (76.6)	15,378 (76.5)	15,522 (76.1)	13,689 (76.0)
Multiple races	22 (0.1)	24 (0.2)	51 (0.3)	46 (0.3)	53 (0.3)	38 (0.2)	41 (0.2)	28 (0.2)
Others	1534 (10.3)	1473 (9.6)	1477 (8.9)	1308 (7.5)	1348 (7.2)	1514 (7.5)	1588 (7.8)	1445 (8.0)
Label, n (%)								
Non-AKI[a]	12,414 (83.1)	12,937 (83.9)	14,097 (84.5)	15,124 (86.7)	16,165 (86.4)	17,435 (86.8)	17,660 (86.6)	15,705 (87.2)
Any AKI	2532 (16.9)	2485 (16.1)	2585 (15.5)	2326 (13.3)	2536 (13.6)	2659 (13.2)	2739 (13.4)	2297 (12.8)
AKI stage ≥2	353 (2.4)	356 (2.3)	359 (2.1)	371 (2.1)	419 (2.2)	471 (2.3)	444 (2.2)	377 (2.1)
AKI stage 3	146 (1.0)	149 (1.0)	171 (1.0)	184 (1.1)	187 (1.0)	241 (1.2)	219 (1.1)	194 (1.1)

[a] AKI: acute kidney injury.

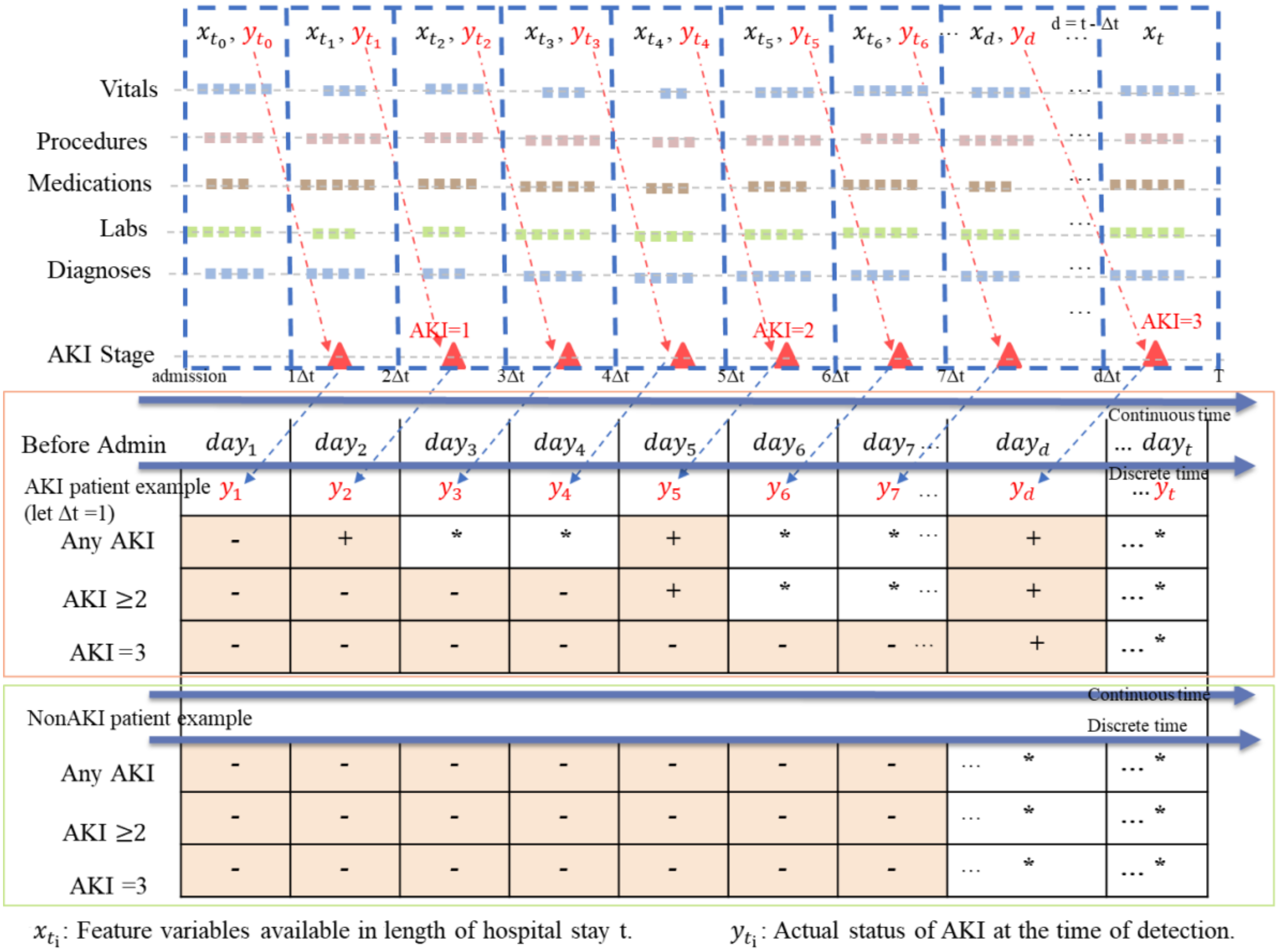
Data Preprocessing

For each patient, we collected all currently populated variables in the PCORNet common data model (CDM) schema, including demographic details (ie, age, gender, and race); structured clinical variables, including comorbidities (International Classification of Diseases-9 and International Classification of Diseases-10 codes), procedures (International Classification of Diseases and Current Procedural Terminology codes), laboratory tests (Logical Observation Identifiers Names and Codes), and medications (RxNorm and National Drug Code); and several vital signs (eg, blood pressure, height, weight, and BMI) [21]. All variables are time stamped, and each sample in the data set is represented by a series of clinical observation vectors aggregated on a daily basis. Therefore, the feature set formed by the data before or on day *t* can be used to predict AKI within days [*t*, *t*+1] for 24-h prediction (or within days [*t*+1, *t*+2] for 48-h prediction).

We preprocessed the data set as follows. First, for numerical features, such as laboratory measurement values and vital signs, we systematically removed the extreme values exceeding 1% and 99%. Second, we performed one-hot coding on categorical variables, such as diagnosis and procedure, to convert them into binary representations. Third, for medication codes, we converted data to cumulative exposure days before the prediction time rather than binary representations. Fourth, the most recent measurement value was chosen when repeated records were available within a certain time interval. Fifth, we used the "sample-and-hold" method to retrieve earlier available measurement values, when measurements were missing for a certain time span. Sixth, we introduced additional features, such as daily blood pressure trend or length of hospital stay, which have been shown to be useful for predicting AKI [22]. Seventh, we excluded all forms of SCr and blood urea nitrogen as they have a high correlation with AKI diagnosis and are not suitable for continuous prediction. Finally, a total of 28,306 features were obtained for model development.

We adopted the discrete-time survival framework [23] to preprocess the time-stamped EHR data, as shown in Figure 1. We divided the patient's entire stay period into *L* nonoverlapping daily windows (ie, *L*=Δ*t*, 2Δ*t*, ..., *T*), where *T* is the length of hospital stay or a specific censor point. Based on expert knowledge, we chose a censor point *T*=7, which represents 7 days since admission. The interval value Δ*t* is the prediction window selected according to clinical needs. For example, Δ*t*=1 means 1-day (24-h) prediction and Δ*t*=2 means 2-day (48-h) prediction. We would use all available data up to time *t*-Δ*t* to predict AKI risk in time *t*. We treated the data corresponding to the AKI-onset day as positive samples based on the criteria of different prediction tasks, while the data after the first positive sample day and between different AKI-stage days were discarded since we could not judge the true AKI stages within these periods because physicians might have intervened and the patient's condition might have improved. All remaining data were regarded as negative samples. For patients who never developed AKI during hospitalization, all available data within 7 days since admission were used to construct negative samples, and other data after 7 days since admission were discarded for the sake of alleviating data imbalance. Under the discrete-time survival framework, we can train a model more in line with real-world clinical practice, where the rolling prediction of AKI risk for a patient on a daily basis is essential [24].

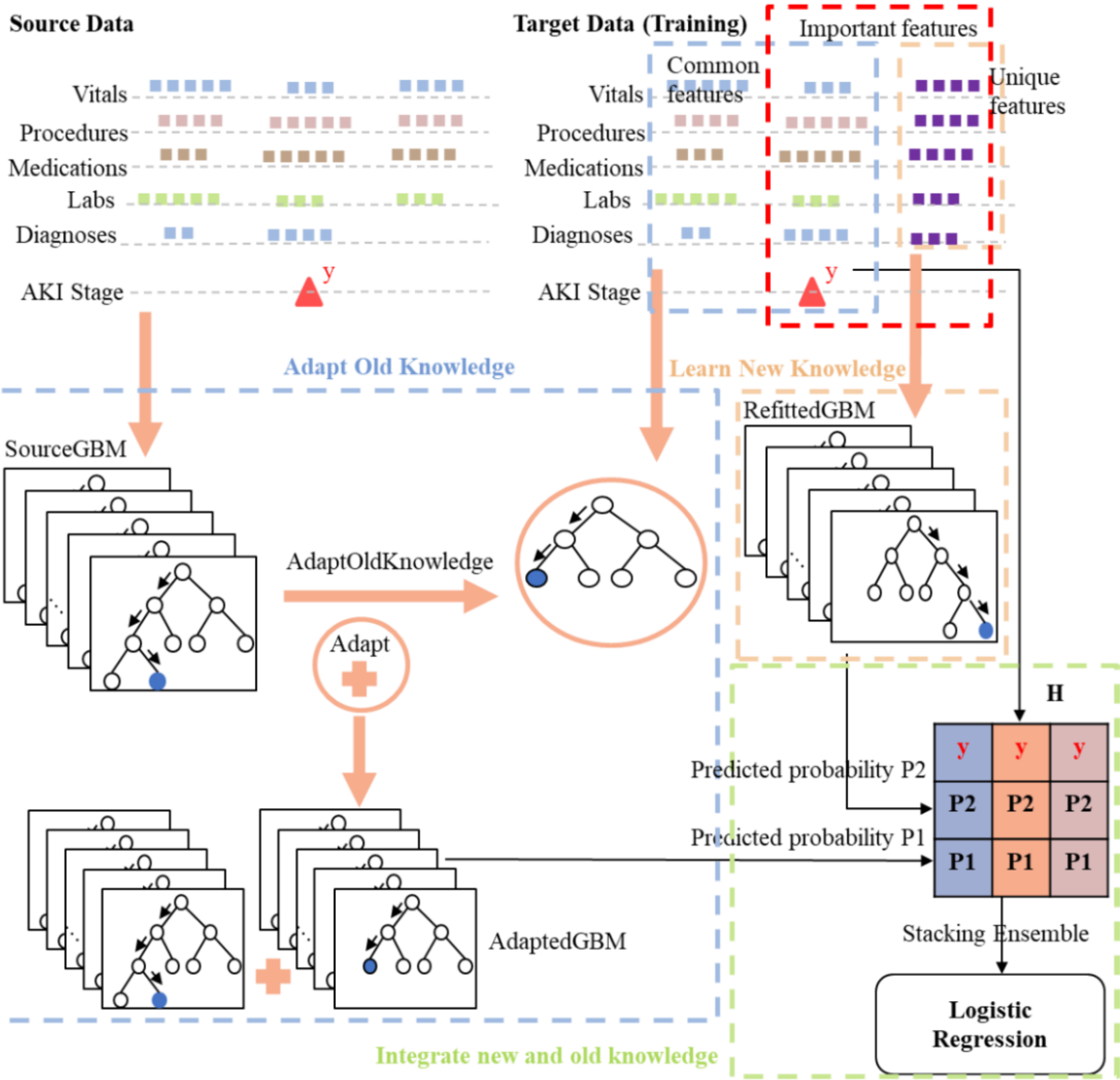
Figure 1. Data processing strategy based on the discrete-time survival framework. The red triangle represents the actual stage of acute kidney injury (AKI). "Δt" indicates the prediction time in advance, "−" indicates negative sample, "+" indicates positive sample, and "*" indicates excluded sample.



TransferGBM Modeling Framework

To correct temporal performance drift, we propose a transfer learning–based modeling framework named TransferGBM, as shown in Figure 2.

Figure 2. Illustration of the TransferGBM modeling framework. AdaptedGBM: adapted gradient boosting machine; AKI: acute kidney injury; RefittedGBM: refitted gradient boosting machine; SourceGBM: source gradient boosting machine; TransferGBM: transfer learning gradient boosting machine.

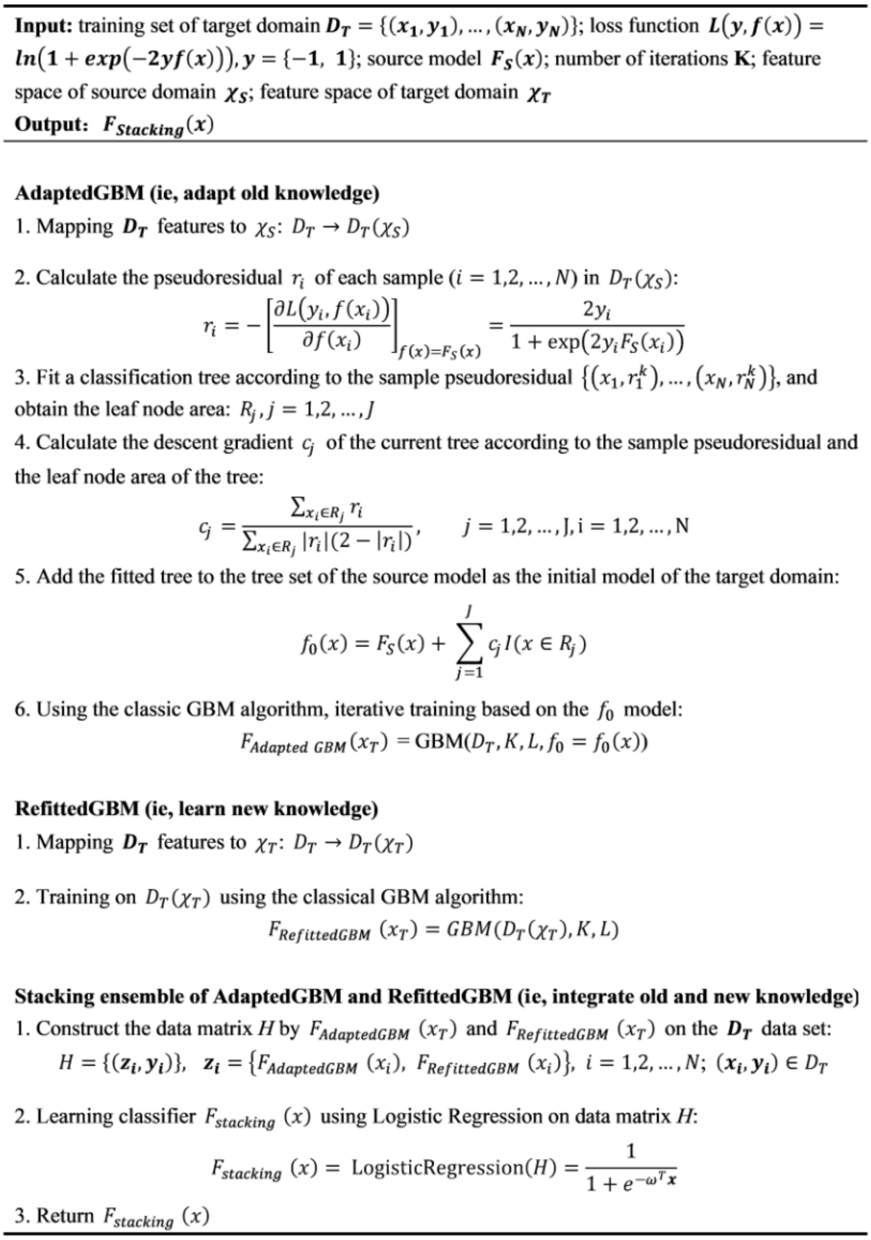


From the perspective of the transfer learning paradigm, we regard the old data as the source domain or source data, and the new data as the target domain or target data. We designed TransferGBM based on several fundamental ideas. First, the base learner is GBM, which has been applied in a wide range of clinical prediction modeling studies [25,26]. GBM has been chosen because (1) it is robust to high-dimensional and collinearity data, (2) it can automatically process missing values, and (3) it embeds a unique feature selection scheme in the model training process, making its output more interpretable [20,27]. Second, we treated the new and old data in different ways, with 2 independent GBM models representing the new and old knowledge, respectively. Third, we transferred old knowledge to the target domain while balancing new and old knowledge in the prediction through an ensemble of the above 2 GBM models. Fourth, we periodically updated the 2 GBM models and their relative weights in the prediction function using target data, in order to adapt to the changing data distribution.

The TransferGBM modeling framework included 5 steps. First, we constructed the source model (ie, source gradient boosting machine [SourceGBM]) using all source data, with a cross-validation–based procedure searching the optimal feature engineering scheme and hyperparameters of GBM (eg, depth of trees, learning rate, minimal child weight, and early stopping). Second, we applied the above optimal feature engineering scheme to the target data and then adapted SourceGBM to the processed target data using the built-in incremental learning mechanism and obtained the adapted model (ie, adapted gradient boosting machine [AdaptedGBM]). Third, we constructed the target model (ie, refitted gradient boosting machine [RefittedGBM]) using the original development set of the target domain while reusing the optimal feature engineering scheme and hyperparameters of GBM from SourceGBM. Fourth, we constructed the predicted probability value matrix for stacking ensemble learning [28], by combining the predicted probability values of AdaptedGBM and RefittedGBM for each sample from the target domain's development set and the true label of the sample into a vector, and pooling all vectors into a matrix *H*. Fifth, we applied the stacking ensemble learning method with the logistic regression (LR) learner to the matrix *H* to obtain the final prediction model, which integrated the old and new knowledge from the AdaptedGBM and RefittedGBM models, respectively.

From the viewpoint of the target domain, the modeling procedure involved 3 distinct sets of features, including (1) the common features that indicate the intersection of the source and target domain features, (2) the unique features that indicate the features belonging to the target domain but not the source domain, and (3) the important features selected by the GBM learner from the target data. When we adapted SourceGBM, we used the common features extracted from the target data combined with missing values of source domain–specific features, so that we could transfer the old knowledge of SourceGBM to the target domain. Considering the value of the target domain–specific knowledge (ie, the new knowledge), we allowed the GBM learner to select the most important features from both the common and unique features of the target data, so that we could obtain the new knowledge of the target domain without constrains on the feature space. The pseudocode of the TransferGBM modeling framework is shown in Figure 3.

Figure 3. Pseudocode of the TransferGBM modeling framework. AdaptedGBM: adapted gradient boosting machine; GBM: gradient boosting machine; RefittedGBM: refitted gradient boosting machine; TransferGBM: transfer learning gradient boosting machine.



```
Input: training set of target domain $D_T = \{(x_1, y_1), \dots, (x_N, y_N)\}$; loss function $L(y, f(x)) = ln(1 + exp(-2yf(x))), y = \{-1,\ 1\}$; source model $F_S(x)$; number of iterations K; feature space of source domain $\chi_S$; feature space of target domain $\chi_T$
Output: $F_{Stacking}(x)$

AdaptedGBM (ie, adapt old knowledge)
1. Mapping $D_T$ features to $\chi_S$: $D_T \rightarrow D_T(\chi_S)$
2. Calculate the pseudoresidual $r_i$ of each sample ($i = 1,2, \dots, N$) in $D_T(\chi_S)$:
       $r_i = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=F_S(x)} = \frac{2y_i}{1 + \exp(2y_i F_S(x_i))}$
3. Fit a classification tree according to the sample pseudoresidual $\{(x_1, r_1^k), \dots, (x_N, r_N^k)\}$, and obtain the leaf node area: $R_j, j = 1,2, \dots, J$
4. Calculate the descent gradient $c_j$ of the current tree according to the sample pseudoresidual and the leaf node area of the tree:
       $c_j = \frac{\sum_{x_i \in R_j} r_i}{\sum_{x_i \in R_j} |r_i|(2 - |r_i|)}, \quad j = 1,2, \dots, \mathrm{J}, \mathrm{i} = 1,2, \dots, \mathrm{N}$
5. Add the fitted tree to the tree set of the source model as the initial model of the target domain:
       $f_0(x) = F_S(x) + \sum_{j=1}^{J} c_j I(x \in R_j)$
6. Using the classic GBM algorithm, iterative training based on the $f_0$ model:
       $F_{Adapted\ GBM}(x_T) = \mathrm{GBM}(D_T, K, L, f_0 = f_0(x))$

RefittedGBM (ie, learn new knowledge)
1. Mapping $D_T$ features to $\chi_T$: $D_T \rightarrow D_T(\chi_T)$
2. Training on $D_T(\chi_T)$ using the classical GBM algorithm:
       $F_{RefittedGBM}(x_T) = GBM(D_T(\chi_T), K, L)$

Stacking ensemble of AdaptedGBM and RefittedGBM (ie, integrate old and new knowledge)
1. Construct the data matrix $H$ by $F_{AdaptedGBM}(x_T)$ and $F_{RefittedGBM}(x_T)$ on the $D_T$ data set:
       $H = \{(z_i, y_i)\}, \ z_i = \{F_{AdaptedGBM}(x_i), \ F_{RefittedGBM}(x_i)\}, \ i = 1,2, \dots, N; \ (x_i, y_i) \in D_T$
2. Learning classifier $F_{stacking}(x)$ using Logistic Regression on data matrix $H$:
       $F_{stacking}(x) = \mathrm{LogisticRegression}(H) = \frac{1}{1 + e^{-\omega^T x}}$
3. Return $F_{stacking}(x)$
```

Experimental Design

We designed the following 3 prediction tasks: any AKI prediction (ie, AKI stage ≥1), moderate-to-severe AKI prediction (ie, AKI stage ≥2), and severe AKI prediction (AKI stage 3). For any AKI prediction, the prediction window was set to 48 hours, while it was 24 hours for the other 2 tasks, according to general clinical needs.

We pooled the 2010 and 2011 data, and used them as old data (ie, a fixed source domain). The data from 2012 to 2017 were used as new data independently, yielding 6 target domains. We applied stratified random sampling to the source and target domain independently, with division into a development set (80%) and a validation set (20%). We tuned the hyperparameters of GBM, including depth of trees (2-10), learning rate (0.01-0.1), minimal child weight (1-10), and number of trees determined by early stopping, on the training set using 10-fold cross-validation. We measured model performance in terms of the AUROC [29], with a mean value from the 95% CI.

It should be noted that the performance of SourceGBM on the target domain's validation set indicated temporal validation and the performance of RefittedGBM (trained using the target domain's development set) on the target domain's validation set indicated internal validation. To validate TransferGBM, we first explored whether there was performance drift over time and then whether TransferGBM could maintain performance.

Ethical Considerations

The study did not require approval from an institutional review board because the data used met the de-identification criteria specified in the Health Insurance Portability and Accountability Act Privacy Rule [30]. The HERON Data Request Oversight Committee approved the data request.

Results

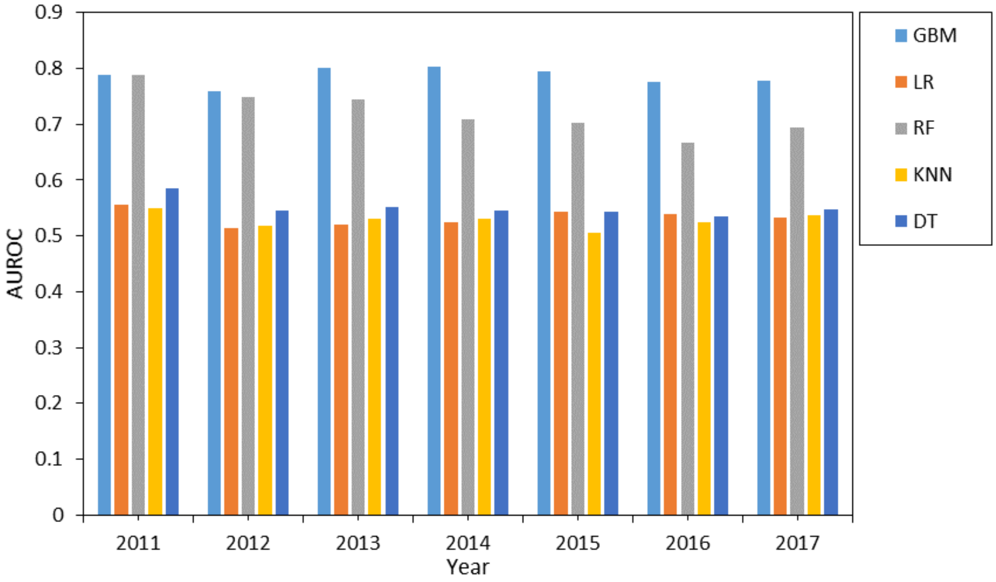
Base Model Selection

We examined 5 common machine learning models based on 5-fold cross-validation on each year's data for any AKI prediction. These models included LR, decision tree (DT), RF, K-nearest neighbor (KNN), and GBM. The model parameters were customized as shown in Table 2, in addition to the default parameters provided in the scikit-learn package [31]. The AUROC performances of the 5 models' internal validations in different years are shown in Figure 4. The AUROCs of both GBM and RF reached 0.7 or above, indicating that these models had a certain predictive ability for AKI, while the performances of the other 3 models (DT, LR, and KNN) were generally poor. Given that GBM performed the best, we chose it as a base learner in the subsequent experiments.

Table 2. Model parameter setting.

Model	Parameter setting (except defaults)
Gradient boosting machine (XG-Boost)	Tune the hyperparameters (depth of trees: 2-10; learning rate: 0.01-0.1; minimal child weight: 1-10) within the development set based on 10-fold cross-validation
Logistic regression	penalty="L2;" max_iter=300; C=3.0
Random forest	n_estimators=400; bootstrap=True
K-nearest neighbor	n_neighbors=40
Decision tree	criterion="entropy"

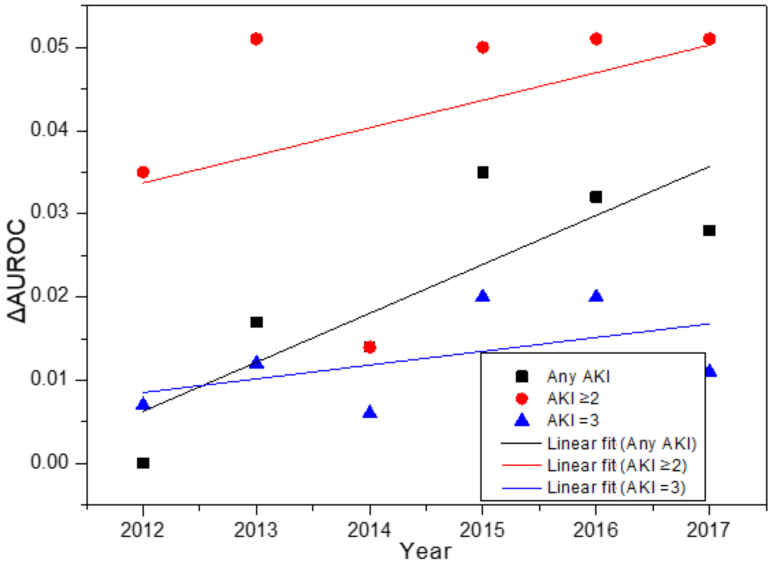
Figure 4. Internal validation of different machine learning models. AUROC: area under the receiver operating characteristic curve; DT: decision tree; GBM: gradient boosting machine; KNN: K-nearest neighbor; LR: logistic regression; RF: random forest.



Performance Shift Over Time

Figure 5 depicts the AUROC gain (ie, ΔAUROC) between the internal validation of RefittedGBM relative to the temporal validation of SourceGBM across 3 prediction tasks. The ΔAUROC shows a linear growth trend over time, implying that the transported model (ie, direct transport of SourceGBM to the target domain without any adaptation) was not the best choice for new data due to the change in data distribution over time. From another point of view, the performance gain was within 0.051, implying that the transported model still contained some general knowledge that can be reused in the new data.

Figure 5. Performance gain by refitting the model. AKI: acute kidney injury; AUROC: area under the receiver operating characteristic curve.



Performance Validation of TransferGBM

TransferGBM maintained a stacking ensemble of 2 GBM models representing new and old knowledge learned from new and old data, respectively, with the former trained using data from 2010 and 2011, and the latter trained using the updated data of each year from 2012 to 2017. Using the validation set of the target domain from 2012 to 2017, we compared model performance between TransferGBM, transported gradient boosting machine (TransportedGBM, ie, direct transport of SourceGBM to the target domain without any adaptation), and RefittedGBM (ie, refitting SourceGBM using the target domain data). To better simulate the process of EHR accumulation in clinical applications, we further investigated different sizes of the available training set (ie, updated data) ranging from 25% to 100% of the target domain's development set via stratified random sampling without replacement. Multimedia Appendix 1 shows the performance in terms of AUROC (95% CI) of TransportedGBM, RefittedGBM, and TransferGBM across different target years and different training set sizes for 3 prediction tasks.

We assessed the impact of different sizes of available training sets on model performance from the perspective of modeling framework selection. Figure 6 illustrates the case of the target year 2012 as an example. The performance of TransportedGBM was better than that of RefittedGBM when the training set size was small. As the amount of training data increased, RefittedGBM gradually improved and finally outperformed TransportedGBM. Overall, regardless of the size of the available training set, the performance of TransferGBM was always better than that of TransportedGBM and RefittedGBM.

Next, we investigated the joint impact of training set size and data distribution shift on model performance regarding the modeling framework selection, as shown in Figure 7.

For any AKI prediction, when the training set was 25%, TransportedGBM outperformed RefittedGBM in the first 3 years (from 2012 to 2014). However, in the subsequent 3 years (from 2015 to 2017), the prediction of TransportedGBM rapidly declined, and it underperformed RefittedGBM. During the whole 6 years, TransferGBM consistently outperformed TransportedGBM and RefittedGBM, with the AUROC ranging from 0.759 (95% CI 0.732-0.766) to 0.804 (95% CI 0.778-0.812), and an average AUROC gain of 0.03 compared to RefittedGBM and 0.02 compared to TransportedGBM. When the training set size was 100%, RefittedGBM significantly outperformed TransportedGBM over all 6 years, but still underperformed TransferGBM. The AUROC of TransferGBM ranged from 0.783 (95% CI 0.757-0.792) to 0.828 (95% CI 0.802-0.834), with an average AUROC gain of 0.04 compared to RefittedGBM and 0.02 compared to TransportedGBM.

For AKI stage ≥2 prediction, even though the training set size was only 25%, RefittedGBM outperformed TransportedGBM (except for target year 2012), and a larger training set was associated with better prediction. This means that the data distribution of the target domain was significantly different from that of the source domain, and directly transporting an external model into the target domain was not a wise choice. Again, TransferGBM was the best model among the 3 models, regardless of the training set size and target year. The AUROC of TransferGBM ranged from 0.830 (95% CI 0.795-0.851) to 0.921 (95% CI 0.893-0.932) when the training set size was 25%, and ranged from 0.866 (95% CI 0.835-0.877) to 0.946 (95% CI 0.920-0.959) when the training set size was 100%.

For AKI stage 3 prediction, when the training set size was 25% or 50%, RefittedGBM significantly underperformed TransportedGBM in the first 3 years (from 2012 to 2014), but the prediction became close in the subsequent 3 years (from 2015 to 2017). When the training set size was 50% or 100%, RefittedGBM and TransportedGBM performed very close to each other. This result implies that direct transportation of an external model was a good choice (ie, there is no need to refit the model, especially when training data on the target domain is not sufficient). TransferGBM was still the best model, and the AUROC ranged from 0.920 (95% CI 0.890-0.936) to 0.948 (95% CI 0.921-0.962) when the training set size was 25%, and ranged from 0.866 (95% CI 0.854-0.911) to 0.959 (95% CI 0.932-0.973) when the training set size was 100%.

Figure 6. Impact of training set size on performance (target year 2012). AKI: acute kidney injury; AUROC: area under the receiver operating characteristic curve; RefittedGBM: refitted gradient boosting machine; TransferGBM: transfer learning gradient boosting machine; TransportedGBM: transported gradient boosting machine.

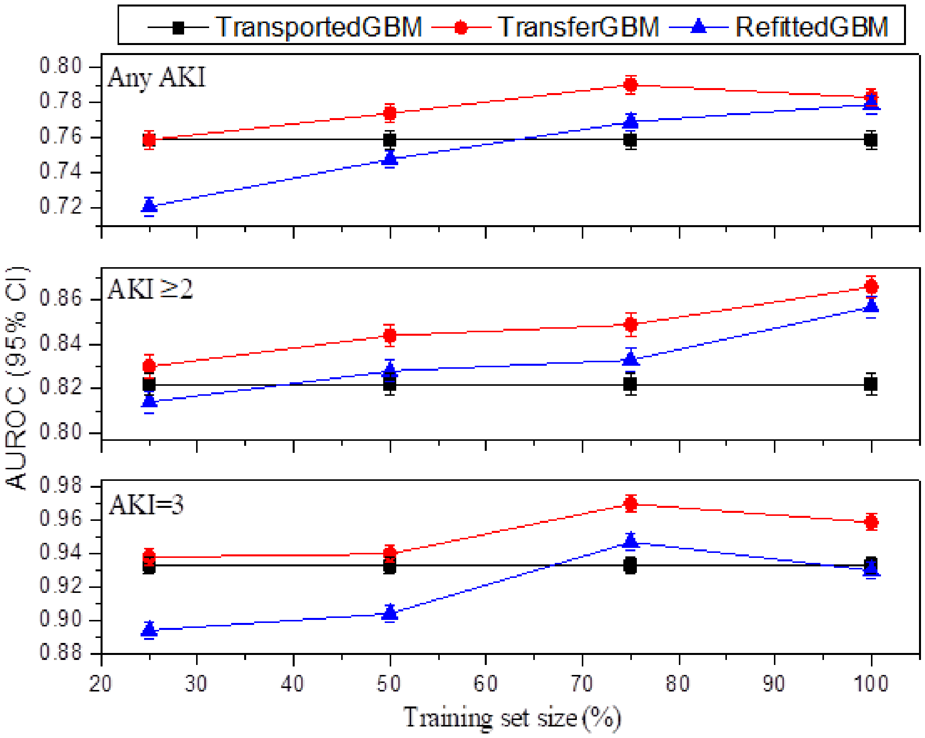
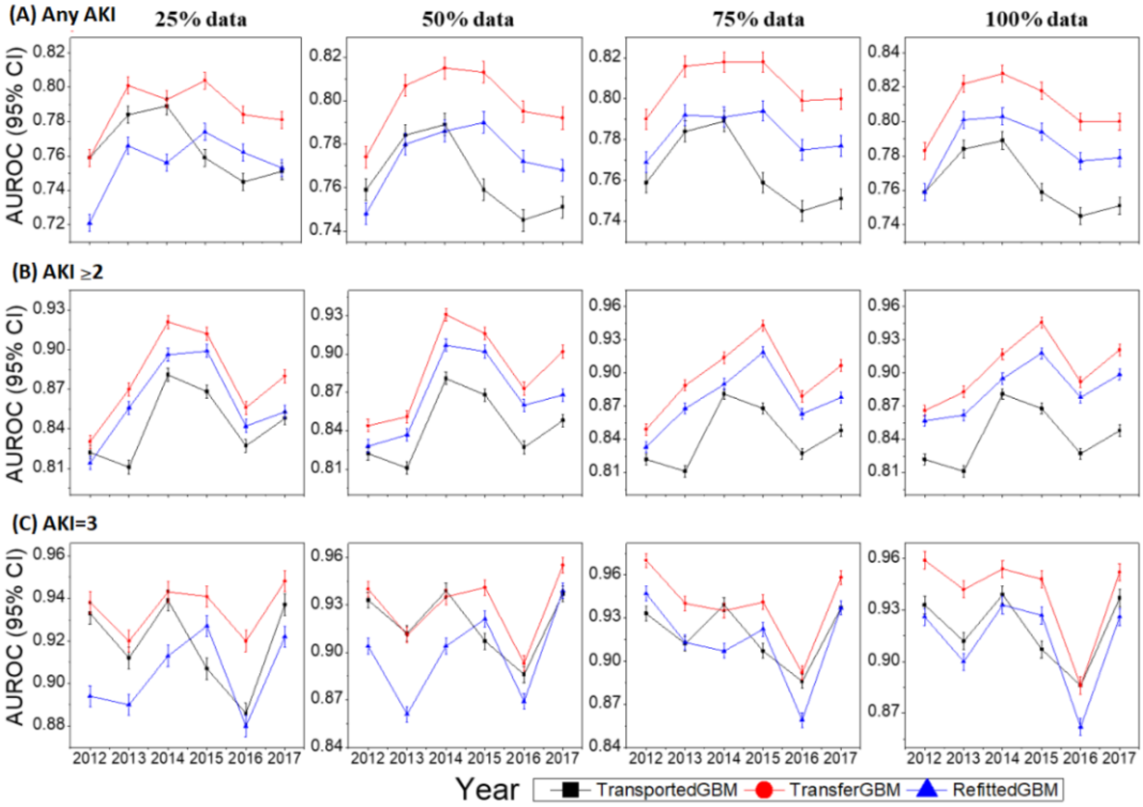


Figure 7. Joint impact of training set size and data distribution shift on performance. AKI: acute kidney injury; AUROC: area under the receiver operating characteristic curve; RefittedGBM: refitted gradient boosting machine; TransferGBM: transfer learning gradient boosting machine; TransportedGBM: transported gradient boosting machine.



Discussion

Principal Findings

Experimental results showed that TransferGBM can consistently outperform TransportedGBM and RefittedGBM, regardless of the amount of available training data from the target domain. We also confirmed that old data are important, and should not be discarded, especially in the case of insufficient new data. There exist differences between old and new knowledge, and thus, there is a need to achieve balance.

With regard to the candidate base learners for the proposed transfer learning–based modeling framework, we considered several commonly used linear and nonlinear machine learning algorithms, and among them, RF has good robustness to overfitting and high-dimensional feature variables [32,33]. XGBoost can consider multiple potentially relevant predictors simultaneously and can handle potentially nonlinear correlations [34-36]. DT is a nonparametric learning algorithm with fast computation and accuracy, can handle continuous and type fields, and is very suitable for high-dimensional data [32]. LR is a linear algorithm that is very suitable for sparse data sets, and the model performance remains stable when only a few variables in the model are valuable predictors. KNN is simple to implement, does not require a data training process, and is very suitable for high-dimensional data. According to the experiment results, the XGBoost algorithm had superior performance. The performance of RF was very close to that of XGBoost, and both were tree-based ensemble approaches. DT may ignore the correlation between variables and experience some large noise, resulting in very poor model performance [33]. The poor performance of LR might be due to the nonlinear correlation between AKI risk factors. KNN may be affected by a large amount of noise in the EHR data, resulting in very poor performance.

The choice of TransportedGBM, RefittedGBM, or TransferGBM depends on or is affected by the actual situation regarding data distribution, modeling cost, available training data from the target domain, etc. TransportedGBM is trained on source data and then is directly applied to the target data without any adaptation and additional cost, which is appropriate for clinical scenarios where the distribution between the source and target domains is very similar. When the distribution is not similar, RefittedGBM would be a better choice than TransportedGBM, and it only requires refitting of the model on the target data, except for the requirement of sufficient training data from the target domain. TransferGBM is no doubt a more complicated solution, which needs to adapt an existing model, refit a new model, and construct an ensemble of these 2 models. This makes TransferGBM more suitable for clinical scenarios where the distribution of the source domain is partially similar to that of the target domain or where the degree of similarity changes significantly.

With regard to the adaptiveness of TransferGBM, it is clear that TransferGBM is a flexible and adaptive extension to the combination of AdaptedGBM and RefittedGBM (AdaptedGBM is obtained by updating TransportedGBM/SourceGBM to the target domain). This also means that TransferGBM might degrade to AdaptedGBM or RefittedGBM due to the stacking ensemble learning mechanism under certain situations. Taking some extreme cases as examples, when the target domain is under the same distribution as the source domain, TransferGBM would degrade to AdaptedGBM and even TransportedGBM since there is little change after updating the model with new data from the target domain. On the contrary, when the target domain is under a distribution completely different from the source domain, TransferGBM would degrade to RefittedGBM, since in this case, AdaptedGBM would be almost useless, and even negative and suppressed under the stacking ensemble learning process. In most cases that TransferGBM is designed for, that is, when the distributions of the source and target domains are more or less similar but not completely different, TransferGBM would adaptively achieve a balance between AdaptedGBM and RefittedGBM.

Motivations

Conventionally, transfer learning is applied to the scenario of data scarcity and distribution disparity, with the underlying idea of selectively reusing data or knowledge from the source domain to assist the modeling process on the target domain. As for the scenario of temporal performance drift, we proposed to regard the old data as the source domain and the new data as the target domain, which might make transfer learning suitable, and we attempted to confirm its effectiveness.

We believe that transfer learning can provide insights from another perspective for correcting temporal performance drift, compared to common approaches such as recalibration and incremental training. For example, when the data distribution significantly changes, transfer learning can immediately discard the old knowledge/model and reselect a new suitable training sample from the source domain to learn, while incremental training suffers from slow progressive adaptation.

Since the primary objective of our study was not to build a high-performance AKI prediction model under the common modeling scenario, we divided the data into different years and adopted a simple and clear modeling process without comprehensive feature engineering, class balancing, hyperparameter searching, etc.

Limitations

There are several limitations associated with our study. First, we used retrospective data in model training and validations, and had not validated our model externally. Thus, our results do not indicate the performance in actual clinical practice. Second, we have not adopted state-of-the-art transfer learning algorithms, such as gapBoost, distant domain transfer learning, selective learning algorithm, multilinear relationship networks, and transitive transfer learning, that have been discussed in systematic reviews [37,38]. These algorithms might yield better prediction performance. Third, we have not compared our method with other correction approaches for temporal performance drift and detection mechanisms of temporal performance drift, such as those proposed by Davis et al [1,2,39]. Fourth, we have not considered prevalent time-series models, such as recurrent neural networks and long short-term memory [40,41], as well as adding historical aggregate feature representations (eg, average laboratory test results and vital signs for the past 48 h) [42]. These methods may yield effects equivalent to those of the transfer learning approach.

Conclusions

This study addressed the problem of performance drift in clinical prediction models. We proposed a novel transfer learning–based modeling framework and validated it using real EHR data from the University of Kansas Medical Center for AKI prediction. The proposed TransferGBM model overcomes the problems of insufficient target data and drifting data distribution through transferring old knowledge and integrating old and new knowledge models. The results showed that TransferGBM is superior to both transported and refitted models.

Acknowledgments

This research was supported by the Major Research Plan of the National Natural Science Foundation of China (Key Program, grant number 91746204), the Science and Technology Development in Guangdong Province (Major Projects of Advanced and Key Techniques Innovation, grant number 2017B030308008), and Guangdong Engineering Technology Research Center for Big Data Precision Healthcare (grant number 603141789047). ML is supported by the National Institutes of Health (NIH)/National Institute of Diabetes and Digestive and Kidney Diseases (award R01DK116986), National Science Foundation Smart and Connected Health (award 2014554), and NIH/National Center for Advancing Translational Sciences (NCATS) Clinical Translational Science Award (CTSA; grant number UL1TR002366). The clinical data set used in this study was obtained from the University of Kansas Medical Center's HERON clinical data repository, which is supported by institutional funding and by the NIH/NCATS CTSA (grant number UL1TR002366).

Authors' Contributions

YH and ML initiated the project and designed the overall study. ML extracted the data used in this study. XZ and KL designed the algorithm. YX, SC, and XS designed the initial training and testing setup, and performed the experiments. YX drafted the paper, with critical revisions by YH, ML, XZ, KL, and WC.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed performance comparison under different experimental settings.
[DOCX File , 30 KB - medinform_v10i11e38053_app1.docx]

References

1. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc 2019 Dec 01;26(12):1448-1457 [FREE Full text] [doi: 10.1093/jamia/ocz127] [Medline: 31397478]
2. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 2017 Nov 01;24(6):1052-1061 [FREE Full text] [doi: 10.1093/jamia/ocx030] [Medline: 28379439]
3. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015 Mar;68(3):279-289 [FREE Full text] [doi: 10.1016/j.jclinepi.2014.06.018] [Medline: 25179855]
4. Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and utility testing of clinical prediction models: Time to change the approach. JAMA 2020 Jul 21;324(3):235-236. [doi: 10.1001/jama.2020.1230] [Medline: 32134437]
5. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012 May;98(9):691-698. [doi: 10.1136/heartjnl-2011-301247] [Medline: 22397946]
6. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009 Jun 04;338:b606. [doi: 10.1136/bmj.b606] [Medline: 19502216]
7. Siregar S, Nieboer D, Vergouwe Y, Versteegh MI, Noyez L, Vonk AB, et al. Improved prediction by dynamic modeling. Circ: Cardiovascular Quality and Outcomes 2016 Mar;9(2):171-181. [doi: 10.1161/circoutcomes.114.001645]
8. Zeng X, McMahon GM, Brunelli SM, Bates DW, Waikar SS. Incidence, outcomes, and comparisons across definitions of AKI in hospitalized individuals. Clin J Am Soc Nephrol 2014 Jan;9(1):12-20 [FREE Full text] [doi: 10.2215/CJN.02730313] [Medline: 24178971]
9. Hoste EAJ, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. Intensive Care Med 2015 Aug;41(8):1411-1423. [doi: 10.1007/s00134-015-3934-7] [Medline: 26162677]

10. Haines RW, Lin S, Hewson R, Kirwan CJ, Torrance HD, O'Dwyer MJ, et al. Acute kidney injury in trauma patients admitted to critical care: Development and validation of a diagnostic prediction model. Sci Rep 2018 Feb 26;8(1):3665 [FREE Full text] [doi: 10.1038/s41598-018-21929-2] [Medline: 29483607]
11. Dai W, Chen Y, Xue G, Yang Q, Yu Y. Translated learning: transfer learning across different feature spaces. In: NIPS'08: Proceedings of the 21st International Conference on Neural Information Processing Systems. 2008 Presented at: 21st International Conference on Neural Information Processing Systems; December 8-10, 2008; Vancouver, British Columbia, Canada p. 353-360. [doi: 10.5555/2981780.2981825]
12. Dai W, Yang Q, Xue G, Yu Y. Boosting for transfer learning. In: ICML '07: Proceedings of the 24th International Conference on Machine Learning. 2007 Presented at: 24th International Conference on Machine Learning; June 20-24, 2007; Corvalis, Oregon, USA p. 193-200. [doi: 10.1145/1273496.1273521]
13. Long M, Wang J, Ding G, Sun J, Yu P. Transfer Feature Learning with Joint Distribution Adaptation. 2013 Presented at: IEEE International Conference on Computer Vision; December 1-8, 2013; Sydney, NSW, Australia. [doi: 10.1109/ICCV.2013.274]
14. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. Knowl. Data Eng 2010 Oct;22(10):1345-1359. [doi: 10.1109/TKDE.2009.191]
15. Segev N, Harel M, Mannor S, Crammer K, El-Yaniv R. Learn on source, refine on target: A model transfer learning framework with random forests. IEEE Trans. Pattern Anal. Mach. Intell 2017 Sep 1;39(9):1811-1824. [doi: 10.1109/tpami.2016.2618118]
16. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016 May 28;3(1):9. [doi: 10.1186/s40537-016-0043-6]
17. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc 2014;21(4):699-706 [FREE Full text] [doi: 10.1136/amiajnl-2013-002162] [Medline: 24481703]
18. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract 2012;120(4):c179-c184 [FREE Full text] [doi: 10.1159/000339789] [Medline: 22890468]
19. Hsu C, Liu C, Tain Y, Kuo C, Lin Y. Machine learning model for risk prediction of community-acquired acute kidney injury hospitalization from electronic health records: Development and validation study. J Med Internet Res 2020 Aug 04;22(8):e16903 [FREE Full text] [doi: 10.2196/16903] [Medline: 32749223]
20. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 2020 Nov 09;11(1):5668 [FREE Full text] [doi: 10.1038/s41467-020-19551-w] [Medline: 33168827]
21. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing knowledge consistently across health systems. Yearb Med Inform 2017 Aug;26(1):139-147 [FREE Full text] [doi: 10.15265/IY-2017-018] [Medline: 29063555]
22. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model. Crit Care Med 2018 Jul;46(7):1070-1077. [doi: 10.1097/CCM.0000000000003123] [Medline: 29596073]
23. Singer JD, Willett JB. It's about time: Using discrete-time survival analysis to study duration and the timing of events. Journal of Educational Statistics 2016 Nov 23;18(2):155-195. [doi: 10.3102/10769986018002155]
24. He J, Hu Y, Zhang X, Wu L, Waitman LR, Liu M. Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. JAMIA Open 2019 Apr;2(1):115-122 [FREE Full text] [doi: 10.1093/jamiaopen/ooy043] [Medline: 30976758]
25. Kim K, Yang H, Yi J, Son H, Ryu J, Kim YC, et al. Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: External validation and model interpretation. J Med Internet Res 2021 Apr 16;23(4):e24120 [FREE Full text] [doi: 10.2196/24120] [Medline: 33861200]
26. Sung M, Hahn S, Han CH, Lee JM, Lee J, Yoo J, et al. Event prediction model considering time and input error using electronic medical records in the intensive care unit: Retrospective study. JMIR Med Inform 2021 Nov 04;9(11):e26426 [FREE Full text] [doi: 10.2196/26426] [Medline: 34734837]
27. Wei C, Zhang L, Feng Y, Ma A, Kang Y. Machine learning model for predicting acute kidney injury progression in critically ill patients. BMC Med Inform Decis Mak 2022 Jan 19;22(1):17 [FREE Full text] [doi: 10.1186/s12911-021-01740-2] [Medline: 35045840]
28. Wolpert DH. Stacked generalization. Neural Networks 1992 Jan;5(2):241-259. [doi: 10.1016/s0893-6080(05)80023-1]
29. Jiménez-Valverde A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography 2012;21(4):498-507. [doi: 10.1111/j.1466-8238.2011.00683.x]
30. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. US Department of Health and Human Services, Office for Human Research Protections. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [accessed 2022-11-03]
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011;12:2825-2830. [doi: 10.5555/1953048.2078195]

32. Corradi JP, Thompson S, Mather JF, Waszynski CM, Dicks RS. Prediction of incident delirium using a random forest classifier. J Med Syst 2018 Nov 14;42(12):261. [doi: 10.1007/s10916-018-1109-0] [Medline: 30430256]
33. Kulkarni VY, Sinha PK, Petare MC. Weighted hybrid decision tree model for random forest classifier. J. Inst. Eng. India Ser. B 2015 Jan 3;97(2):209-217. [doi: 10.1007/s40031-014-0176-y]
34. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. XGBoost: eXtreme Gradient Boosting, R package version 04-2. R Project. 2015. URL: https://cran.r-project.org/src/contrib/Archive/xgboost/ [accessed 2022-10-19]
35. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: 10.1145/2939672.2939785]
36. Ogunleye A, Wang Q. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform 2020;17(6):2131-2140. [doi: 10.1109/TCBB.2019.2911071] [Medline: 30998478]
37. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A comprehensive survey on transfer learning. Proc. IEEE 2021 Jan;109(1):43-76. [doi: 10.1109/jproc.2020.3004555]
38. Niu S, Liu Y, Wang J, Song H. A decade survey of transfer learning (2010–2020). IEEE Trans. Artif. Intell 2020 Oct;1(2):151-166. [doi: 10.1109/tai.2021.3054609]
39. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. J Biomed Inform 2020 Dec;112:103611 [FREE Full text] [doi: 10.1016/j.jbi.2020.103611] [Medline: 33157313]
40. 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]
41. Yadav P, Steinbach M, Kumar V, Simon G. Mining Electronic Health Records (EHRs). ACM Comput. Surv 2018 Nov 30;50(6):85. [doi: 10.1145/3127881]
42. 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 31;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]

Abbreviations

AdaptedGBM: adapted gradient boosting machine
AKI: acute kidney injury
AUROC: area under the receiver operating characteristic curve
DT: decision tree
EHR: electronic health record
GBM: gradient boosting machine
KNN: K-nearest neighbor
LR: logistic regression
RefittedGBM: refitted gradient boosting machine
RF: random forest
SCr: serum creatinine
SourceGBM: source gradient boosting machine
TransferGBM: transfer learning gradient boosting machine
TransportedGBM: transported gradient boosting machine

Edited by T Hao; submitted 17.03.22; peer-reviewed by H Li, K Zheng, H Sun; comments to author 15.05.22; revised version received 31.07.22; accepted 12.10.22; published 09.11.22.

Please cite as:
Zhang X, Xue Y, Su X, Chen S, Liu K, Chen W, Liu M, Hu Y
A Transfer Learning Approach to Correct the Temporal Performance Drift of Clinical Prediction Models: Retrospective Cohort Study
JMIR Med Inform 2022;10(11):e38053
URL: https://medinform.jmir.org/2022/11/e38053
doi:10.2196/38053
PMID:36350705

Original Paper

Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning–Based Text Simplification Approach

Atharva Phatak[1], MSc; David W Savage[2], MD, PhD; Robert Ohle[3], MSc, MA, MBBCh; Jonathan Smith[2], MD; Vijay Mago[1], PhD

[1]Department of Computer Science, Lakehead University, Thunder Bay, ON, Canada
[2]NOSM University, Thunder Bay, ON, Canada
[3]NOSM University, Sudbury, ON, Canada

Corresponding Author:
Atharva Phatak, MSc
Department of Computer Science
Lakehead University
955 Oliver Road
Thunder Bay, ON, P7B 5E1
Canada
Phone: 1 8073558351
Email: phataka@lakeheadu.ca

Abstract

Background: In most cases, the abstracts of articles in the medical domain are publicly available. Although these are accessible by everyone, they are hard to comprehend for a wider audience due to the complex medical vocabulary. Thus, simplifying these complex abstracts is essential to make medical research accessible to the general public.

Objective: This study aims to develop a deep learning–based text simplification (TS) approach that converts complex medical text into a simpler version while maintaining the quality of the generated text.

Methods: A TS approach using reinforcement learning and transformer–based language models was developed. Relevance reward, Flesch-Kincaid reward, and lexical simplicity reward were optimized to help simplify jargon-dense complex medical paragraphs to their simpler versions while retaining the quality of the text. The model was trained using 3568 complex-simple medical paragraphs and evaluated on 480 paragraphs via the help of automated metrics and human annotation.

Results: The proposed method outperformed previous baselines on Flesch-Kincaid scores (11.84) and achieved comparable performance with other baselines when measured using ROUGE-1 (0.39), ROUGE-2 (0.11), and SARI scores (0.40). Manual evaluation showed that percentage agreement between human annotators was more than 70% when factors such as fluency, coherence, and adequacy were considered.

Conclusions: A unique medical TS approach is successfully developed that leverages reinforcement learning and accurately simplifies complex medical paragraphs, thereby increasing their readability. The proposed TS approach can be applied to automatically generate simplified text for complex medical text data, which would enhance the accessibility of biomedical research to a wider audience.

(JMIR Med Inform 2022;10(11):e38095) doi:10.2196/38095

KEYWORDS

medical text simplification; reinforcement learning; natural language processing; manual evaluation

Introduction

Background

Research from the field of biomedicine contains essential information about new clinical trials on topics related to new drugs and treatments for a variety of diseases. Although this information is publicly available, it often has complex medical terminology, making it difficult for the general public to understand. One way to address this problem is by converting the complex medical text into a simpler language that can be understood by a wider audience. Although manual text simplification (TS) is one way to address the problem, it cannot be scaled to the rapidly expanding body of biomedical literature.

Therefore, there is a need for the development of *natural language processing* approaches that can automatically perform TS.

Related Studies

TS Approaches

Initial research in the field of TS focused on *lexical simplification* (LS) [1,2]. An LS system typically involves replacing complex words with their simpler alternatives using lexical databases, such as the *Paraphrase Database* [3], WordNet [4], or using language models, such as *bidirectional encoder representations from transformer*s (BERT) [5]. Recent research defines TS as a *sequence-to-sequence* (seq2seq) task and has approached it by leveraging model architectures from other seq2seq tasks such as machine translation and text summarization [6-8]. Nisioi et al [9] proposed a neural *seq2seq* model, which used *long short-term memories* (LSTMs) for automatic TS. It was trained on simple-complex sentence pairs and showed through human evaluations that the TS system–generated outputs ultimately preserved meaning and were grammatically correct [9]. Afzal et al [10] incorporated LSTMs to create a quality-aware text summarization system for medical data. Zhang and Lapata [11] developed an LSTM-based neural encoder-decoder TS model and trained it using *reinforcement learning* (RL) to directly optimize SARI [12] scores along with a few other rewards. SARI is a widely used metric for automatic evaluation of TS.

With the recent progress in natural language processing research, LSTM-based models were outperformed by transformer [13]-based language models [13-16]. Transformers follow an encoder-decoder structure with both the encoder and decoder made up of *L* identical layers. Each layer consists of 2 sublayers, one being a feed-forward layer and the other a multihead attention layer. Transformer-based language models, such as BART [14], generative pretraining transformer (GPT) [15], and *text-to-text-transfer-transformer* [16], have achieved strong performance on natural language generation tasks such as text summarization and machine translation.

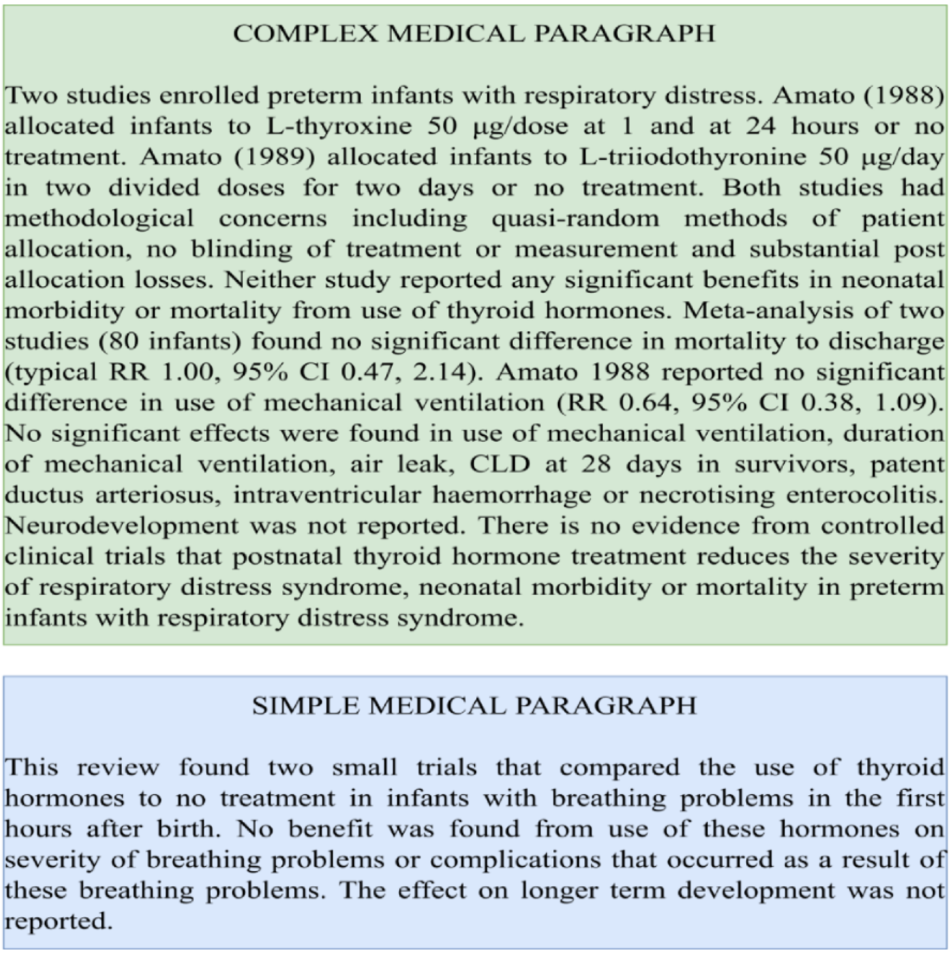
Building on the success of transformer-based language models, recently Martin et al [17] introduced *multilingual unsupervised sentence simplification* (MUSS) [17], a BART [14]-based language model, which achieved state-of-the-art performance on TS benchmarks by training on paraphrases mined from CCNet [18] corpus. Zhao et al [19] proposed a semisupervised approach that incorporated the back-translation architecture along with denoising autoencoders for the purpose of automatic TS. Unsupervised TS is also an active area of research but has been primarily limited to LS. However, in a recent study, Surya et al [20] proposed an unsupervised approach to perform TS at both the lexical and syntactic levels. In general, research in the field of TS has been focused mostly on sentence-level simplification. However, Sun et al [21] proposed a document-level data set (D-wikipedia) and baseline models to perform document-level simplification. Similarly, Devaraj et al [8] proposed a BART [14]-based model that was trained using unlikelihood loss for the purpose of paragraph-level medical TS. Although their training regime penalizes the terms considered “jargon” and increases the readability, the generated text has lower quality and diversity [8]. Thus, the lack of document- or paragraph-level simplification makes this an important work in the advancement of the field.

TS Data Sets

The majority of TS research uses data extracted from Wikipedia and news articles [11,22,23]. These data sets are paired sentence-level data sets (ie, for each complex sentence, there is a corresponding simple sentence). TS systems have heavily relied on sentence-level data sets, extracted from regular and simple English Wikipedia, such as WikiLarge [11], because they are publicly available. It was later shown by Xu [24] that there are issues with data quality for the data sets extracted from Wikipedia. They proposed the Newsela corpus, which was created by educators who rewrote news articles for different school-grade levels. Automatic sentence alignment methods [25] were used on the Newsela corpus to create a sentence-level TS data set. Despite the advancements in research on sentence-level simplification, there is a need for TS systems that can simplify text at a paragraph level.

Recent work has focused on the construction of document-level simplification data sets [17,21,26]. Sun et al [21] constructed a document-level data set, called D-Wikipedia, by aligning the English Wikipedia and Simple English Wikipedia spanning 143,546 article pairs. Although there are many data sets available for sentence-level TS, data sets for domain-specific paragraph-level TS are lacking. In the field of medical TS, Van den Bercken et al [27] constructed a sentence-level simplification data set using sentence alignment methods. Recently, Devaraj et al [8] proposed the first paragraph-level medical simplification data set, containing 4459 simple-complex pairs of text, and this is the data set used for the analysis and baseline training in this study. A snippet of a complex paragraph and its simplified version from the data set proposed by Devaraj et al [8] is shown in Figure 1. The data set is open sourced and publicly available [28].

Figure 1. Complex medical paragraph and the corresponding simple medical paragraph from the dataset.



COMPLEX MEDICAL PARAGRAPH

Two studies enrolled preterm infants with respiratory distress. Amato (1988) allocated infants to L-thyroxine 50 µg/dose at 1 and at 24 hours or no treatment. Amato (1989) allocated infants to L-triiodothyronine 50 µg/day in two divided doses for two days or no treatment. Both studies had methodological concerns including quasi-random methods of patient allocation, no blinding of treatment or measurement and substantial post allocation losses. Neither study reported any significant benefits in neonatal morbidity or mortality from use of thyroid hormones. Meta-analysis of two studies (80 infants) found no significant difference in mortality to discharge (typical RR 1.00, 95% CI 0.47, 2.14). Amato 1988 reported no significant difference in use of mechanical ventilation (RR 0.64, 95% CI 0.38, 1.09). No significant effects were found in use of mechanical ventilation, duration of mechanical ventilation, air leak, CLD at 28 days in survivors, patent ductus arteriosus, intraventricular haemorrhage or necrotising enterocolitis. Neurodevelopment was not reported. There is no evidence from controlled clinical trials that postnatal thyroid hormone treatment reduces the severity of respiratory distress syndrome, neonatal morbidity or mortality in preterm infants with respiratory distress syndrome.

SIMPLE MEDICAL PARAGRAPH

This review found two small trials that compared the use of thyroid hormones to no treatment in infants with breathing problems in the first hours after birth. No benefit was found from use of these hormones on severity of breathing problems or complications that occurred as a result of these breathing problems. The effect on longer term development was not reported.

TS Evaluation

The evaluation of TS usually falls into 2 categories: automatic evaluations and manual (ie, human) evaluations. Because of the subjective nature of TS, it has been suggested that the best approach is to perform manual evaluations, based on criteria such as fluency, meaning preservation, and simplicity [20]. Automatic evaluation metrics most commonly used include readability indices such as Flesch-Kincaid Reading Ease [29], *Flesch-Kincaid Grade Level* (FKGL) [29], *Automated Readability Index* (ARI), Coleman-Liau index, and metrics for natural language generation tasks such as SARI [12] and BLEU [30].

Readability indices are used to assign a grade level to text signifying its simplicity. All the readability indices are calculated using some combination of word weighting, syllable, letter, or word counts, and are shown to measure some level of simplicity. Automatic evaluation metrics, such as BLEU [30] and SARI [12], are widely used in TS research, with SARI [12] having specifically been developed for TS tasks. SARI is computed by comparing the generated simplifications with both the source and target references. It computes an average of F_1-score for 3 *n-gram* overlap operations: additions, keeps, and deletions. Both BLEU [30] and SARI [12] are n-gram–based metrics, which may fail to capture the semantics of the generated text.

Objective

The aim of this study is to develop an automatic TS approach that is capable of simplifying medical text data at a paragraph level, with the goal of providing greater accessibility of biomedical research. This paper uses RL-based training to directly optimize 2 properties of simplified text: relevance and simplicity. *Relevance* is defined as simplified text that retains salient and semantic information from the original article. *Simplicity* is defined as simplified text that is easy to understand and lexically simple. These 2 properties are optimized using TS-specific rewards, resulting in a system that outperforms previous baselines on Flesch-Kincaid scores. Extensive human evaluations are conducted with the help of domain experts to judge the quality of the generated text.

The remainder of the paper is organized as follows: The “Methods” section provides details on the data set, the training procedure, and the proposed model, and describes how automatic and human evaluations were conducted to analyze the outputs generated by the proposed model (TESLEA). The “Results” section provides a brief description of the baseline models and the results obtained by conducting automatic and manual evaluation of the generated text. Finally under the “Discussion” section, we highlight the limitations, future work, and draw conclusions.

Methods

Model Objective

Given a complex medical paragraph, the goal of this work is to generate a simplified paragraph that is concise and captures the salient information expressed in the complex text. To accomplish this, an RL-based simplification model is proposed, which optimizes multiple rewards during training, and is tuned using a paragraph-level medical TS data set.

Data Set

The Cochrane Database of Scientific Reviews is a health care database with information on a wide range of clinical topics. Each review includes a plain language summary (PLS) written by the authors who follow guidelines to structure the summaries. PLSs are supposed to be clear, understandable, and accessible, especially for a general audience not familiar with the field of medicine. PLSs are highly heterogeneous in nature, and are not paired (ie, for every complex sentence there may not be a corresponding simpler version). However, Devaraj et al [8] used the Cochrane Database of Scientific Reviews data to produce a paired data set, which has 4459 pairs of complex-simple text, with each text containing less than 1024 tokens so that it can be fed into the BART [14] model for the purpose of TS. The pioneering data set developed by Devaraj et al [8] is used in this study for training the models and is publicly available [28].

TESLEA: TS Using RL

Model and Rewards

The TS solution proposed for the task of simplifying complex medical text uses an RL-based simplification model, which optimizes multiple rewards (*relevance reward*, *Flesch-Kincaid Grade rewards, and lexical simplicity rewards*) to achieve a more complete and concise simplification. The following subsections introduce the computation of these rewards, along with the training procedure.

Relevance Reward

Relevance reward measures how well the semantics of the target text is captured in its simplified version. This is calculated by computing the cosine similarity between the target text embedding (E_T) and the generated text embedding (E_G). BioSentVec [31], a text embedding model trained on medical documents, is used to generate the text embeddings. The steps to calculate the relevance score are depicted in Algorithm 1.

The *RelevanceReward* function takes 3 arguments as input, namely, target text (T), generated text (G), and the embedding model (M). The function *ComputeEmbedding* takes the input text and embedding model (M) as input and generates the relevant text embedding. Finally, cosine similarity between generated text embedding (E_G) and target text embedding (E_T) is calculated to get the reward (Algorithm 1, line 4).

Flesch-Kincaid Grade Reward

FKGL refers to the grade level that must be attained to comprehend the presented information. A higher FKGL score indicates that the text is more complex, and a lower score indicates that the text is simpler. The FKGL for a text (S) is calculated using equation 1 [29]:

$$\text{FKGL(S)} = 0.38 \times (\text{total words/total sentences}) + 1.8 \times (\text{total syllables/total words}) - (15.59) \quad \mathbf{(1)}$$

The FKGL reward (R_{Flesch}) is designed to reduce the complexity of generated text and is calculated as presented in Algorithm 2.

In Algorithm 2, the function *FleschKincaidReward* takes 2 arguments as inputs, namely, generated text (G) and target text (T). The *FKGLScore* function calculates the FKGL for the given text. Once the FKGL for T and G is calculated, the Flesch-Kincaid reward (R_{Flesch}) is calculated as the relative difference between $r(T)$ and $r(G)$ (Algorithm 2, line 4), where $r(T)$ and $r(G)$ denote the FKGL of the target and generated text.

Lexical Simplicity Reward

Lexical simplicity is used to measure whether the words in the generated text (G) are simpler than the words in the source text (S). Laban et al [26] proposed a lexical simplicity reward that uses the correlation between word difficulty and word frequency [32]. As word frequency follows *zipf law*, Laban et al [26] used it to design the reward function, which involves calculating *zipf* frequency of newly inserted words, that is, $Z(G – S)$, and deleted words, that is, $Z(S – G)$. The lexical simplicity reward is defined in the same way as proposed by Laban et al [26] and is described in Algorithm 3. The analysis of the data set proposed by Devaraj et al [8] revealed that 87% of simple and complex pairs have a value of $\Delta Z(S, G) \approx 0.4$, where $\Delta Z(S, G) = Z(G – S) – Z(S – G)$ is the difference between the *zipf* frequency of inserted words and deleted words, with the value of lexical reward ($R_{lexical}$) scaled between 0 and 1.

In Algorithm 3, *LexicalSimplicityReward* requires the source text (S) and the generated text (G) as the inputs. Functions *ZIPFInserted* [25] and *ZIPFDeleted* [25] calculate the *zipf* frequency of newly inserted words and the deleted words. Finally, the lexical reward ($R_{lexical}$) is calculated and normalized, as described in line 5.

Training Procedure and Baseline Model

Pretrained BART

The baseline language model used in this study for performing simplification was BART [14], which is a transformer based encoder-decoder model that was pretrained using a denoising objective function. The decoder part of the model is autoregressive in nature, making it more suitable for sentence-generation tasks. Furthermore, the BART model achieves strong performance on natural language generation tasks such as summarization, which was demonstrated on XSum

[33] and CNN/Daily Mail [34] data sets. In this case, a version of BART fine-tuned on XSUM [33] data set is being used.

Language Model Fine-tuning

Transformer-based language models are pretrained on a large corpus of text and later fine-tuned on a downstream task by minimizing the maximum likelihood loss (*Lml*) function [3]. Consider a paired data set *C*, where each instance consists of a source sentence containing *n* tokens $x = \{x_1,\ldots,x_n\}$ and target sequence containing *m* tokens $y = \{y_1,\ldots,y_n\}$, then the *Lml* function is given in equation 2 with the computation described in Algorithm 4.

where θ represents the model parameters and $y_{<t}$ denotes preceding tokens before the position *t* [35].

However, the results obtained by minimizing *Lml* are not always optimal. There are 2 main reasons for the degradation of results. The first is called "exposure bias" [36], which occurs when the model expects gold-standard data at each step of training, but does not receive appropriate supervision during testing, resulting in an accumulation of errors during prediction. The second is called "representation collapse" [37], which is a degradation of the pretrained language model representations during fine-tuning. Ranzato et al [36] avoided the problem of exposure bias by directly optimizing the specific discrete metric instead of minimizing the *Lml* with the help of an RL-based algorithm called REINFORCE [38]. A variant of REINFORCE [38] called Self-Critical Sequence Training [39] was used in this study to directly optimize certain rewards specifically designed for TS; more information on this is provided in the following subsection.

Self-critical Sequence Training

TS can be formulated as an RL problem, where the "agent" (language model) interacts with the environment to take "action" (next word prediction) based on a learned "policy" (p_θ) defined by model parameters θwhile observing some rewards (*R*). In this work, BART [14] was used as the language model, and the REINFORCE [38] algorithm was used to learn an optimal policy that maximizes rewards. Specifically, REINFORCE was used with a baseline to stabilize the training procedure using an objective function (*Lpg*) with a baseline reward *b* (equation 3):

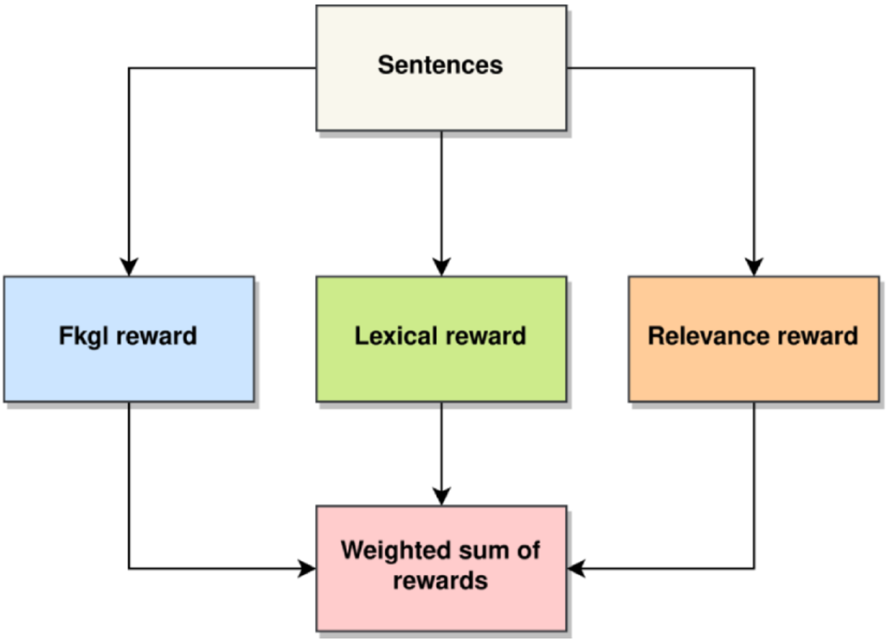
where $p_\theta(y_i^s|\ldots)$ denotes the probability of the *i*th word conditioned on a previously generated sampled sequence by the model; $r(y^s)$ denotes the reward computed for a sentence generated using sampling; denotes the source sentence, and *n* is the length of the generated sentence. Rewards are computed as a weighted sum of the relevance reward (R_{cosine}), R_{Flesch}, and lexical simplicity reward ($R_{lexical}$; Figure 2) and are given by:

where α, β, and *d* are the weights associated with the rewards, respectively.

To approximate the baseline reward, Self-Critical Sequence Training [39] was used. The baseline was calculated by computing reward values for a sentence that has been generated using greedy decoding $r(y^*)$ by the current model and its computation is described in Algorithm 5. The loss function is defined in equation 5:

where y^* denotes the sentence generated using greedy decoding. More details on greedy decoding are described in Multimedia Appendix 1 (see also [8,14,17,25,26,39-42]).

Figure 2. Compute Rewards function calculates a weighted sum of three rewards: Fkgl Reward, Lexical Simplicity Reward, Relevance Reward.



Intuitively, by minimizing the loss described in equation 5, the likelihood of choosing the samples sequence (y^s) is promoted if the reward obtained for sampled sequence, $r(y^s)$, is greater than the reward obtained for the baseline rewards, that is, the samples that return higher reward than $r(y^*)$. The samples that obtain a lower reward are subsequently suppressed. The model is trained using a combination of *Lml* and policy gradient loss similar to [43]. The overall loss is given as follows:

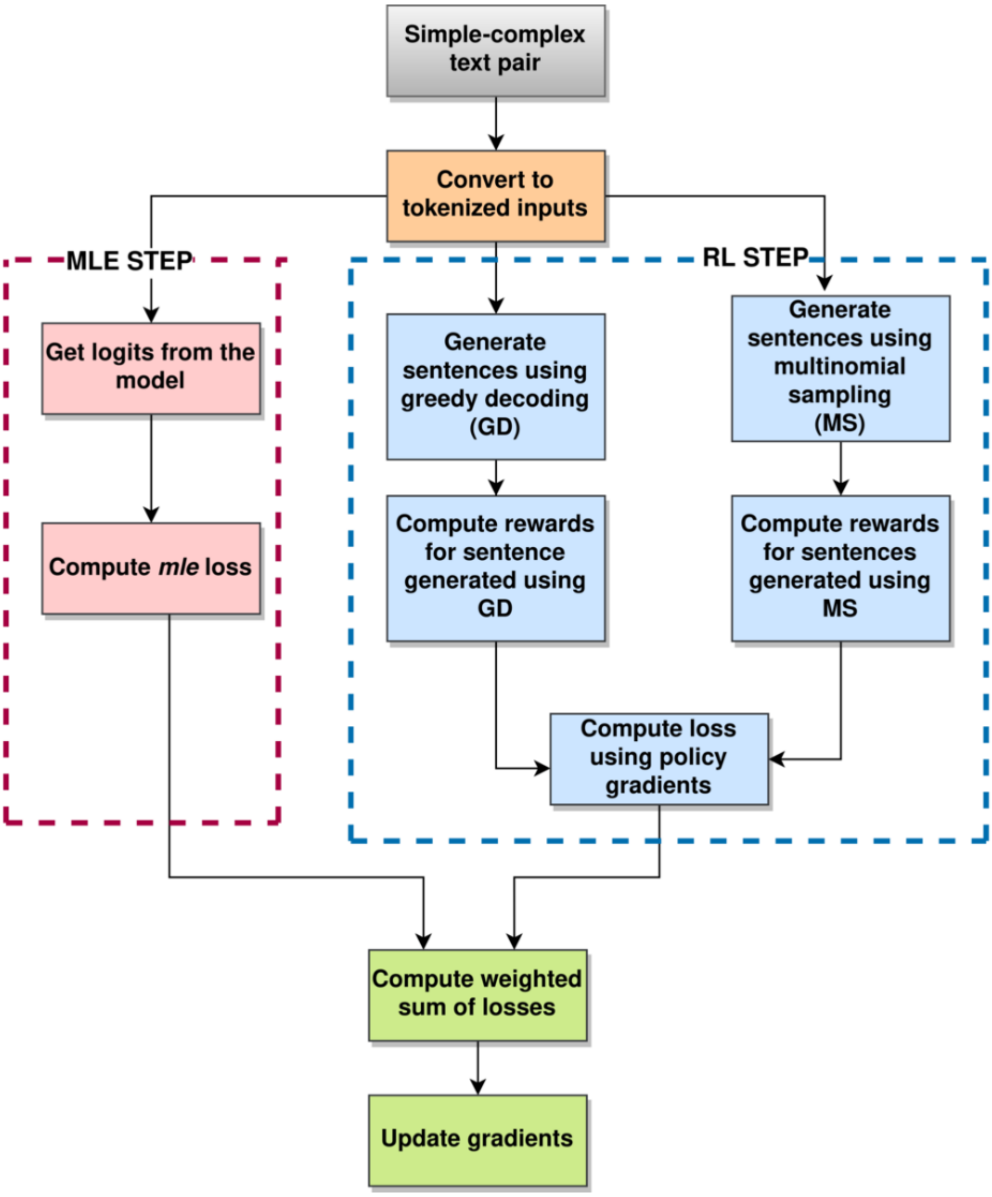
$$L = \gamma Lpg + (1 - \gamma)Lml \quad \mathbf{(6)}$$

where γ is a scaling factor that can be tuned.

Summary of the Training Process

Overall, the training procedure follows a 2-step approach. As the pretrained BART [14] was not trained on the medical domain–related text, it was first fine-tuned on the document-level paired data set [8] by minimizing the *Lml* (maximum likelihood estimation [MLE]; equation 2). In the second part, the fine-tuned BART model was trained further using RL. The RL procedure of TESLEA involves 2 steps: (1) the RL step and (2) the MLE optimization step, which are both shown in Figure 3 and further described in Algorithm 6. The given simple-complex text pairs are converted to tokens as required by the BART model. In the MLE step, these tokens are used to compute *logits* from the model, and then finally MLE loss is computed. In the RL step, the model generates simplified text using 2 decoding strategies: (1) greedy decoding and (2) multinomial sampling. Rewards are computed as weighted sums (Figure 3) for sentences generated using both the decoding strategies. These rewards are then used to calculate the loss for the RL step. Finally, a weighted sum of losses is computed that is used to estimate the gradients and update model parameters. All the hyperparameter settings used are included in Multimedia Appendix 2 (see also [8,12,29,33,34,44-47]).

Figure 3. Reinforcement learning–based training procedure for TESLEA. MLE: maximum likelihood estimation; RL: reinforcement learning.


Simple-complex text pair
Convert to tokenized inputs
MLE STEP
RL STEP
Get logits from the model
Compute *mle* loss
Generate sentences using greedy decoding (GD)
Generate sentences using multinomial sampling (MS)
Compute rewards for sentence generated using GD
Compute rewards for sentences generated using MS
Compute loss using policy gradients
Compute weighted sum of losses
Update gradients

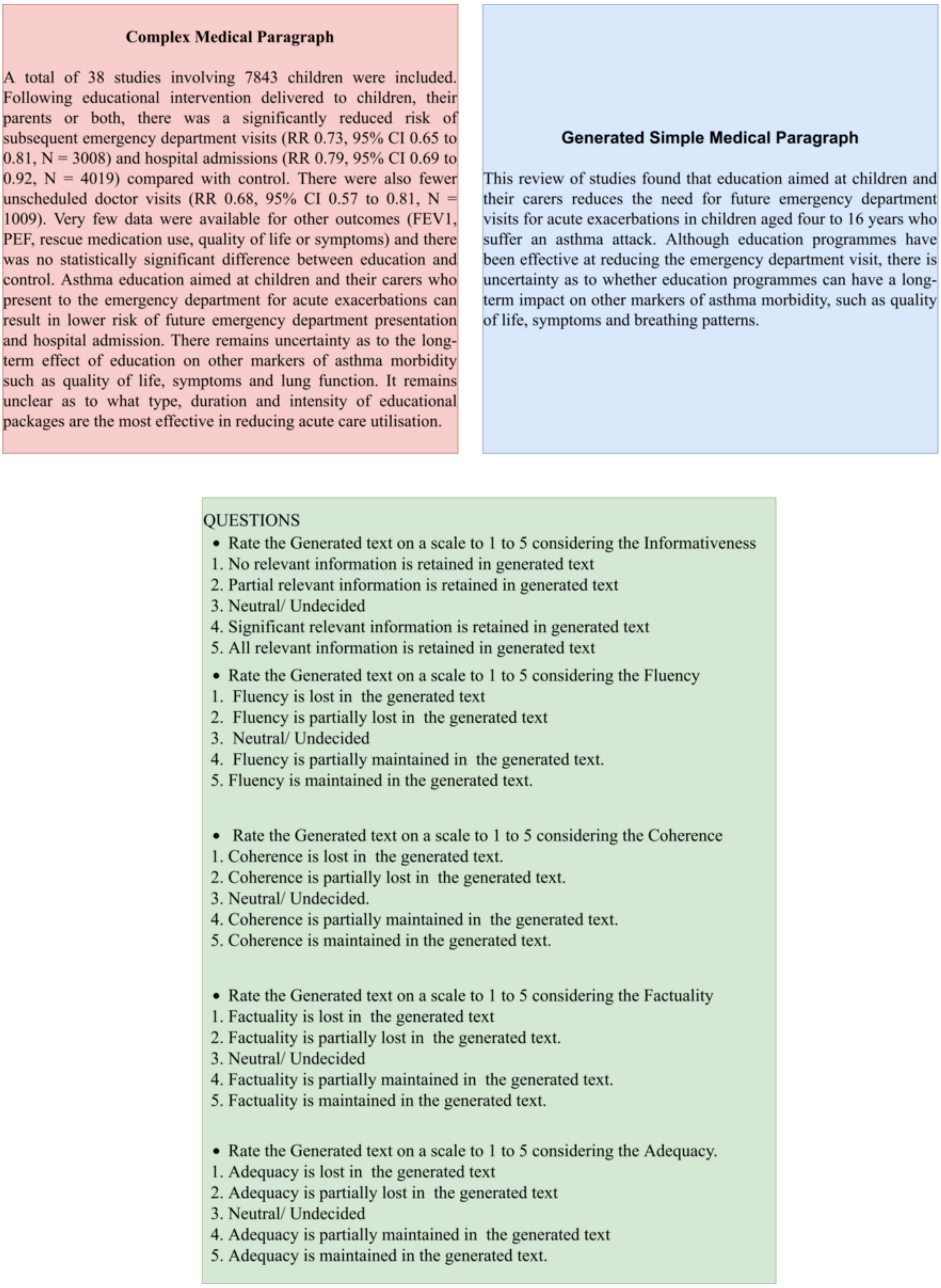
Automatic Metrics

Two readability indices were used to perform automatic evaluations of the generated text, namely, FKGL and Automatic Readability Indices (ARIs). The SARI score is a standard metric for TS. The F-1 versions of ROUGE-1 and ROUGE-2 [44] scores were also reported. Readers can find more details about these metrics in Multimedia Appendix 2. To measure the quality of the generated text, the criteria proposed by Yuan et al [45] were used, which are mentioned in the "Automatic Evaluation Metrics" section in Multimedia Appendix 2. The criteria proposed by Yuan et al [45] can be automatically computed using a language model–based metric called "BARTScore." Further details on how to use BARTScore to measure the quality of the generated text are also mentioned in Multimedia Appendix 2.

Human Evaluations

In this study, 3-domain experts judge the quality of the generated text based on the factors mentioned in the previous section. The evaluators rate the text on a Likert scale from 1 to 5. First, simplified test data were generated using TESLEA, and then 51 generated paragraphs were randomly selected, creating 3 subsets containing 17 paragraphs each. Every evaluator was presented with 2 subsets, that is, a total of 34 complex-simple TESLEA-generated paragraphs. The evaluations were conducted via Google Forms, and the human annotators were asked to measure the quality of simplification for informativeness (INFO), fluency (FLU), coherence (COH), factuality (FAC), and adequacy (ADE) (Figure 4). All the data collected were stored in CSV files for statistical analysis.

Figure 4. A sample question seen by the human annotator.



Complex Medical Paragraph

A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019) compared with control. There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.

Generated Simple Medical Paragraph

This review of studies found that education aimed at children and their carers reduces the need for future emergency department visits for acute exacerbations in children aged four to 16 years who suffer an asthma attack. Although education programmes have been effective at reducing the emergency department visit, there is uncertainty as to whether education programmes can have a long-term impact on other markers of asthma morbidity, such as quality of life, symptoms and breathing patterns.

QUESTIONS

- Rate the Generated text on a scale to 1 to 5 considering the Informativeness

1. No relevant information is retained in generated text
2. Partial relevant information is retained in generated text
3. Neutral/ Undecided
4. Significant relevant information is retained in generated text
5. All relevant information is retained in generated text

- Rate the Generated text on a scale to 1 to 5 considering the Fluency

1. Fluency is lost in the generated text
2. Fluency is partially lost in the generated text
3. Neutral/ Undecided
4. Fluency is partially maintained in the generated text.
5. Fluency is maintained in the generated text.

- Rate the Generated text on a scale to 1 to 5 considering the Coherence

1. Coherence is lost in the generated text.
2. Coherence is partially lost in the generated text.
3. Neutral/ Undecided.
4. Coherence is partially maintained in the generated text.
5. Coherence is maintained in the generated text.

- Rate the Generated text on a scale to 1 to 5 considering the Factuality

1. Factuality is lost in the generated text
2. Factuality is partially lost in the generated text.
3. Neutral/ Undecided
4. Factuality is partially maintained in the generated text.
5. Factuality is maintained in the generated text.

- Rate the Generated text on a scale to 1 to 5 considering the Adequacy.

1. Adequacy is lost in the generated text
2. Adequacy is partially lost in the generated text
3. Neutral/ Undecided
4. Adequacy is partially maintained in the generated text
5. Adequacy is maintained in the generated text.

Results

Overview

This section consists of 3 subsections, namely, (1) Baseline Models, (2) Automatic Evaluations, and (3) Human Evaluations. The first section highlights the baseline models used for comparison and analysis. The second section discusses the results obtained by performing automatic evaluations of the model. The third and final section discusses the results obtained from human assessments and analyzes the relationship between human annotations and automatic metrics.

Baseline Models

TESLEA is compared with other strong baseline models and their details are discussed below:

- BART-Fine-tuned: BART-Fine-tuned is a BART-large model fine-tuned using an *Lml* on the data set proposed by

Devaraj et al [8]. Studies have shown that large pretrained models often perform competitively when fine-tuned for downstream tasks, thus making this a strong competitor.

- BART-UL: Devaraj et al [8] also proposed BART-UL for paragraph-level medical TS. It is the first model to perform paragraph-level medical TS and has achieved strong results on automated metrics. BART-UL was trained using an unlikelihood objective function that penalizes the model for generating technical words (ie, complex words). Further details on the training procedure of BART-UL are described in Multimedia Appendix 1.
- MUSS: MUSS [17] is a BART-based language model that was trained by mining paraphrases from the CCNet corpus [18]. MUSS was trained on a data set consisting of 1 million paraphrases, helping it achieve a strong SARI score. Although MUSS is trained on a sentence-level data set, it still serves as a strong baseline for comparison. Further details on the training procedure for MUSS are discussed in Multimedia Appendix 1

.

- Keep it Simple (KIS): Laban et al [26] proposed an unsupervised approach for paragraph-level TS. KIS is trained using RL and uses the GPT-2 model as a backbone. KIS has shown strong performance on SARI scores beating many supervised and unsupervised TS approaches. Additional details on the training procedure for KIS are described in Multimedia Appendix 1.
- PEGASUS models: PEGASUS is a transformer-based encoder-decoder model that has achieved state-of-the-art results on many text-summarization data sets. It was specifically designed for the task of text summarization. In our analysis, we used 2 variants of PEGASUS models, namely, (1) PEGASUS-large, the large variant of Pegasus model, and (2) PEGASUS-pubmed-large, the large variant of the PEGASUS model that was pretrained on a PubMed data set. Both the PEGASUS models were fine-tuned using *Lml* on the data set proposed by Devaraj et al [8]. For more information regarding the PEGASUS model, the readers are suggested to refer to [46].

The models described above are the only ones available for medical TS as of June 2022.

Results of Automatic Metrics

The metrics used for automatic evaluation are FKGL, ARI, ROUGE-1, ROUGE-2, SARI, and BARTScore. The mean readability indices scores (ie, FKGL and ARI) obtained by various models are reported in Table 1. ROUGE-1, ROUGE-2, and SARI scores are reported in Table 2 and BARTScore is reported in Table 3.

Table 1. Flesch-Kincaid Grade Level and Automatic Readability Index for the generated text.[a]

Text	Flesch-Kincaid Grade Level	Automatic Readability Index
Baseline		
Technical abstracts	14.42	15.58
Gold-standard references	13.11	15.08
Model generated		
BART-Fine-tuned	13.45	15.32
BART-UL	11.97	13.73[b]
TESLEA	11.84[b]	13.82
MUSS[c]	14.29	17.29
Keep it Simple	14.15	17.05
PEGASUS-large	14.53	17.55
PEGASUS-pubmed-large	16.35	19.8

[a] TESLEA significantly reduces FKGL and ARI scores when compared with plain language summaries.

[b] Best score.

[c] MUSS: multilingual unsupervised sentence simplification.

Table 2. ROUGE-1, ROUGE-2, and SARI scores for the generated text.[a]

Model	ROUGE-1	ROUGE-2	SARI
BART-Fine-tuned	0.40	0.11	0.39
BART-UL	0.38	0.14	0.40[b]
TESLEA	0.39	0.11	0.40[b]
MUSS[c]	0.23	0.03	0.34
Keep it Simple	0.23	0.03	0.32
PEGASUS-large	0.44[b]	0.18[b]	0.40[b]
PEGASUS-pubmed-large	0.42	0.16	0.40[b]

[a] TESLEA achieves similar performance to other models. Higher scores of ROUGE-1, ROUGE-2, and SARI are desirable.

[b] Best performance.

[c] MUSS: multilingual unsupervised sentence simplification.

Table 3. Faithfulness Score and F-score for the generated text by the models.[a]

Models	Faithfulness Score	F-score
BART-Fine-tuned	0.137	0.078
BART-UL	0.242	0.061
TESLEA	0.366[b]	0.097[b]
MUSS[c]	0.031	0.029
Keep it Simple	0.030	0.028
PEGASUS-large	0.197	0.073
PEGASUS-pubmed-large	0.29	0.063

[a] Higher scores of Faithfulness and F-score are desirable.

[b] Highest score.

[c] MUSS: multilingual unsupervised sentence simplification.

Readability Indices, ROUGE, and SARI Scores

The readability indices scores reported in Table 1 suggest that the FKGL scores obtained by TESLEA are better (ie, a lower score) when compared with the FKGL scores obtained by comparing technical abstracts (ie, complex medical paragraphs available in the data set) with the gold-standard references (ie, simple medical paragraphs corresponding to the complex medical paragraphs). Moreover, TESLEA achieves the lowest FKGL score (11.84) when compared with baseline models, indicating significant improvement in the TS. The results suggest that (1) BART-based transformer models are capable of performing simplification at the paragraph level such that the outputs are at a reduced reading level (FKGL) when compared with technical abstracts, gold-standard references, and baseline models. (2) The proposed method to optimize TS-specific rewards allows the generation of text with greater readability than even the gold-standard references, as indicated by the FKGL scores in Table 1. The reduction in FKGL scores can be explained by the fact that FKGL was a part of a reward (R_{Flesch}) that was directly being optimized.

In addition, we report the SARI [12] and ROUGE scores [44] as shown in Table 2. SARI is a standard automatic metric used in sentence-level TS tasks. The ROUGE score is another standard metric in text summarization tasks. The results show that TESLEA matches the performance of baseline models on both ROUGE and SARI scores. Although there are no clear patterns when ROUGE and SARI scores are considered, there are differences in the quality of text generated by these models and these are explained in the “Text Quality Measure” subsection.

Text Quality Measure

There has been significant progress in designing automatic metrics that are able to capture linguistic quality of the text generated by language models. One such metric that is able to measure the quality of generated text is BARTScore [45]. BARTScore has shown strong correlation with human assessments on various tasks ranging from machine translation to text summarization. BARTScore has 4 different metrics (ie, Faithfulness Score, Precision, Recall, F-score), which can be used to measure different qualities of generated text. Further details on how to use BARTScore are mentioned in Multimedia Appendix 2.

According to the analysis conducted by Yuan et al [45], Faithfulness Score measures 3 aspects of generated text via COH, FLU, and FAC. The F-score measures 2 aspects of generated text (INFO and ADE). In our analysis, we use these

2 variants of BARTScore to measure COH, FLU, FAC, INFO, and ADE. TESLEA achieves the highest values (Table 3) of Faithfulness Score (0.366) and F-score (0.097), indicating that the rewards designed for the purpose of TS not only help the model in generating simplified text but also on some level preserve the quality of generated text. The F-scores of all the models are relatively poor (ie, scores closer to 1 are desirable). One of the reasons for low F-scores could be the introduction of misinformation or hallucinations in the generated text, a common problem for language models, which could be addressed by adapting training strategies that focus on INFO via the help of rewards or objective functions.

For qualitative analysis we randomly selected 50 sentences from the test data and calculated the average number of tokens based on BART model vocabulary. For the readability measure, we calculated the FKGL scores of these generated texts and noted any textual inconsistencies such as misinformation. The analysis revealed that the text generated by most models was significantly smaller than the gold-standard references (Table 4). Furthermore, TESLEA- and BART-UL–generated texts were significantly shorter compared with other baseline models and TESLEA had the lowest FKGL score among all the models as depicted in Table 4.

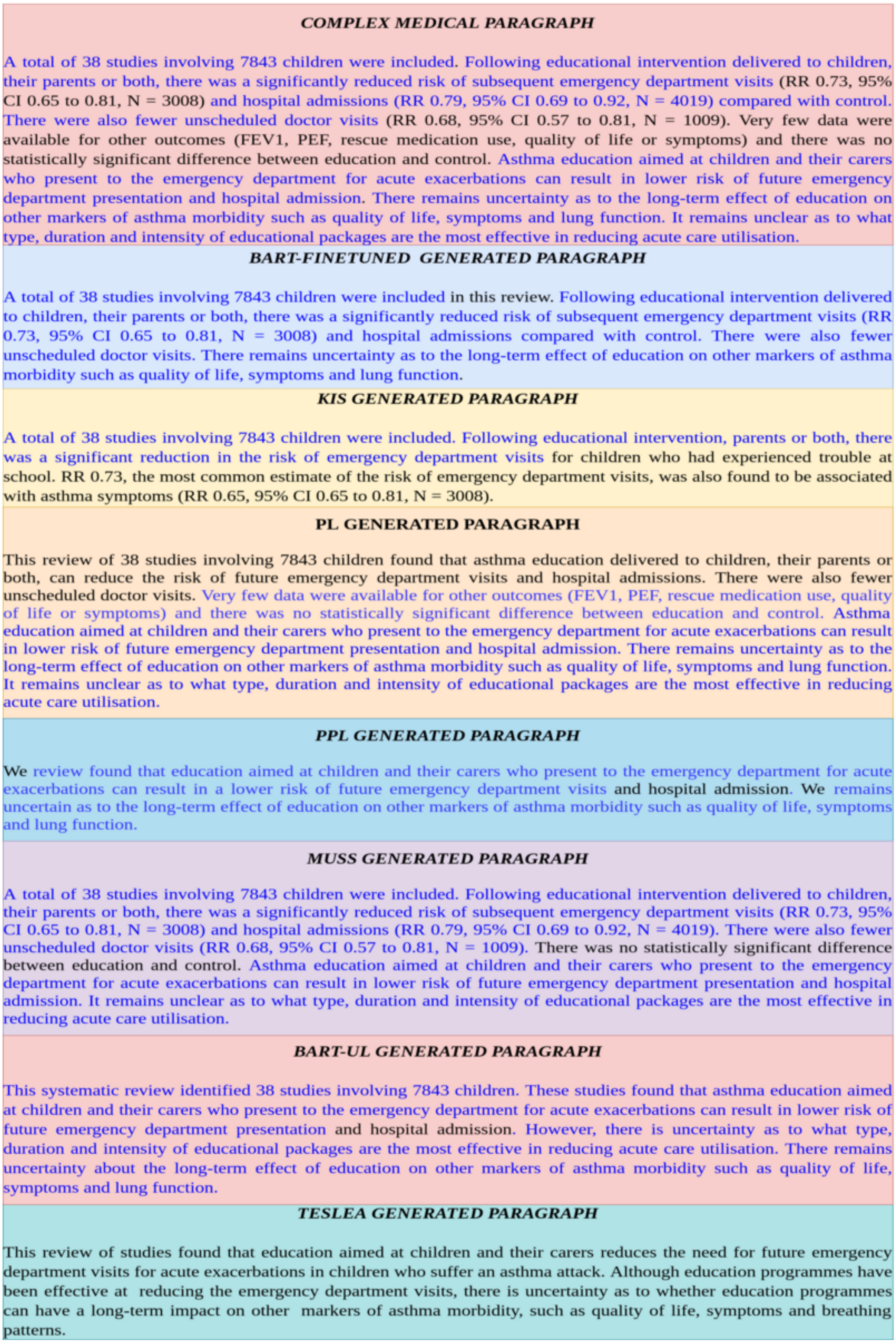
From a qualitative point of view, the sentences generated by most baseline models involve significant duplication of text from the original complex medical paragraph. The outputs generated by the KIS model were incomplete and appear "noisy" in nature. One of the reasons for the noise generation could be because of unstable training due to lack of a huge corpus of domain-specific data. BART-UL–generated paragraphs are simplified as indicated by the FKGL and ARI scores, but they are extractive in nature (ie, the model learns to select simplified sentences from the original medical paragraph and combines them to form a simplification). PEGASUS-pubmed-large–generated paragraphs are also extractive in nature and similar to BART-UL–generated paragraphs, but it was observed that they were grammatically inconsistent. In contrast to baseline models, the text generated by TESLEA was concise, semantically relevant, and simple, without involving any medical domain–related complex vocabulary. Figure 5 shows an example of text generated by all the models, with blue text indicating the copied text.

In addition to the duplicated text, the models also induced misinformation in the generated text. The most common form of induced misinformation observed was "The evidence is current up to [date]," as shown in Figure 6. This text error occurred due to the structure of the data (ie, PLS contains statements related to this research, but these statements were not in the original text; thus, the model attempted to add these statements to the generated text although it is not factually correct). Thus considerable attention should be paid to including FAC measures in the training regime of these models. For a more complete assessment of the quality of simplification, human evaluation was conducted using domain experts for the text generated by TESLEA.

Table 4. Average number of tokens and average Flesch-Kincaid Grade Level scores for selected samples.

Model	Number of tokens	Flesch-Kincaid Grade Level
Technical abstracts	498.11	14.37
Gold-standard references	269.74	12.77
TESLEA	131.37	12.34
BART-UL	145.08	12.66
Keep it Simple	187.59	13.78
Multilingual unsupervised sentence simplification	193.07	13.86
PEGASUS-large	272.04	13.93
PEGASUS-pubmed-large	150.00	15.09

Figure 5. Comparison of Text Generated by all the models. The highlighted blue text indicates copying. CI: Confidence Interval; FEV: Force Expiratory Volume; N: Population size; PEV: Peak Expiratory Flow; RR: Respiratory Rate.



COMPLEX MEDICAL PARAGRAPH

A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019) compared with control. There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.

BART-FINETUNED GENERATED PARAGRAPH

A total of 38 studies involving 7843 children were included in this review. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions compared with control. There were also fewer unscheduled doctor visits. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function.

KIS GENERATED PARAGRAPH

A total of 38 studies involving 7843 children were included. Following educational intervention, parents or both, there was a significant reduction in the risk of emergency department visits for children who had experienced trouble at school. RR 0.73, the most common estimate of the risk of emergency department visits, was also found to be associated with asthma symptoms (RR 0.65, 95% CI 0.65 to 0.81, N = 3008).

PL GENERATED PARAGRAPH

This review of 38 studies involving 7843 children found that asthma education delivered to children, their parents or both, can reduce the risk of future emergency department visits and hospital admissions. There were also fewer unscheduled doctor visits. Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.

PPL GENERATED PARAGRAPH

We review found that education aimed at children and their carers who present to the emergency department for acute exacerbations can result in a lower risk of future emergency department visits and hospital admission. We remains uncertain as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function.

MUSS GENERATED PARAGRAPH

A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019). There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). There was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.

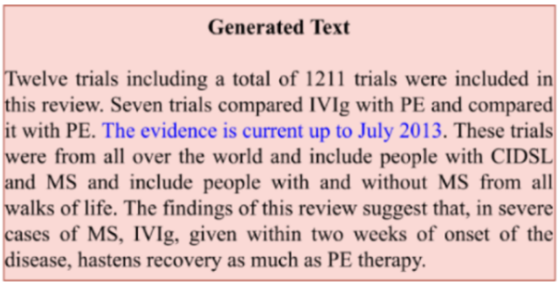
BART-UL GENERATED PARAGRAPH

This systematic review identified 38 studies involving 7843 children. These studies found that asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. However, there is uncertainty as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation. There remains uncertainty about the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function.

TESLEA GENERATED PARAGRAPH

This review of studies found that education aimed at children and their carers reduces the need for future emergency department visits for acute exacerbations in children who suffer an asthma attack. Although education programmes have been effective at reducing the emergency department visits, there is uncertainty as to whether education programmes can have a long-term impact on other markers of asthma morbidity, such as quality of life, symptoms and breathing patterns.

Figure 6. Example of misinformation found in Generated text. CIDSL: Cornelia de Lange syndrome; IVIg: Intravenous immune globulin; MS: Multiple Sclerosis; PE: plasma exchange.



Generated Text

Twelve trials including a total of 1211 trials were included in this review. Seven trials compared IVIg with PE and compared it with PE. The evidence is current up to July 2013. These trials were from all over the world and include people with CIDSL and MS and include people with and without MS from all walks of life. The findings of this review suggest that, in severe cases of MS, IVIg, given within two weeks of onset of the disease, hastens recovery as much as PE therapy.

Human Evaluations

For this research, 3 domain experts assessed the quality of generated text, based on factors of INFO, FLU, COH, FAC, and ADE, as proposed by Yuan et al [45], which are discussed in Multimedia Appendix 2. To measure interrater reliability, the percentage agreement between the annotators is calculated, and the results are shown in Table 5. The average percentage agreement for the factors of FLU, COH, FAC, and ADE is the highest, indicating that annotators agree among their evaluations.

The average Likert score for each factor is also reported by each rater (Table 6). From the data mentioned in Table 6, the raters think that the COH and FLU have the highest quality, with the ADE, FAC, and INFO also rated reasonably high.

To further assess whether results obtained by automated metrics truly signify an improvement in the quality of generated text by TESLEA, the Spearman rank correlation coefficient was calculated between human ratings and the automatic metrics for all 51 generated paragraphs (text), with the results shown in Table 7. The BARTScore has the highest correlation with human ratings for FLU, FAC, COH, and ADE compared with other metrics. A few text samples along with their human annotations and automated metric scores are shown in Multimedia Appendix 3 and Figure 7.

Table 5. Average percentage interrater agreement.

Interrater agreement	Informativeness, %	Fluency, %	Factuality, %	Coherence, %	Adequacy, %
A1[a] and A2[b]	82.35	82.35	82.35	70.59	82.35
A1 and A3[c]	70.59	58.82	70.59	70.59	70.59
A3 and A2	52.94	70.59	74.51	74.51	64.71
Average (% agreement)	68.63	70.59	74.51	74.51	72.55

[a] A1: annotator 1.

[b] A2: annotator 2.

[c] A3: annotator 3.

Table 6. Average Likert score by each rater for informativeness, fluency, factuality, coherence, and adequacy.

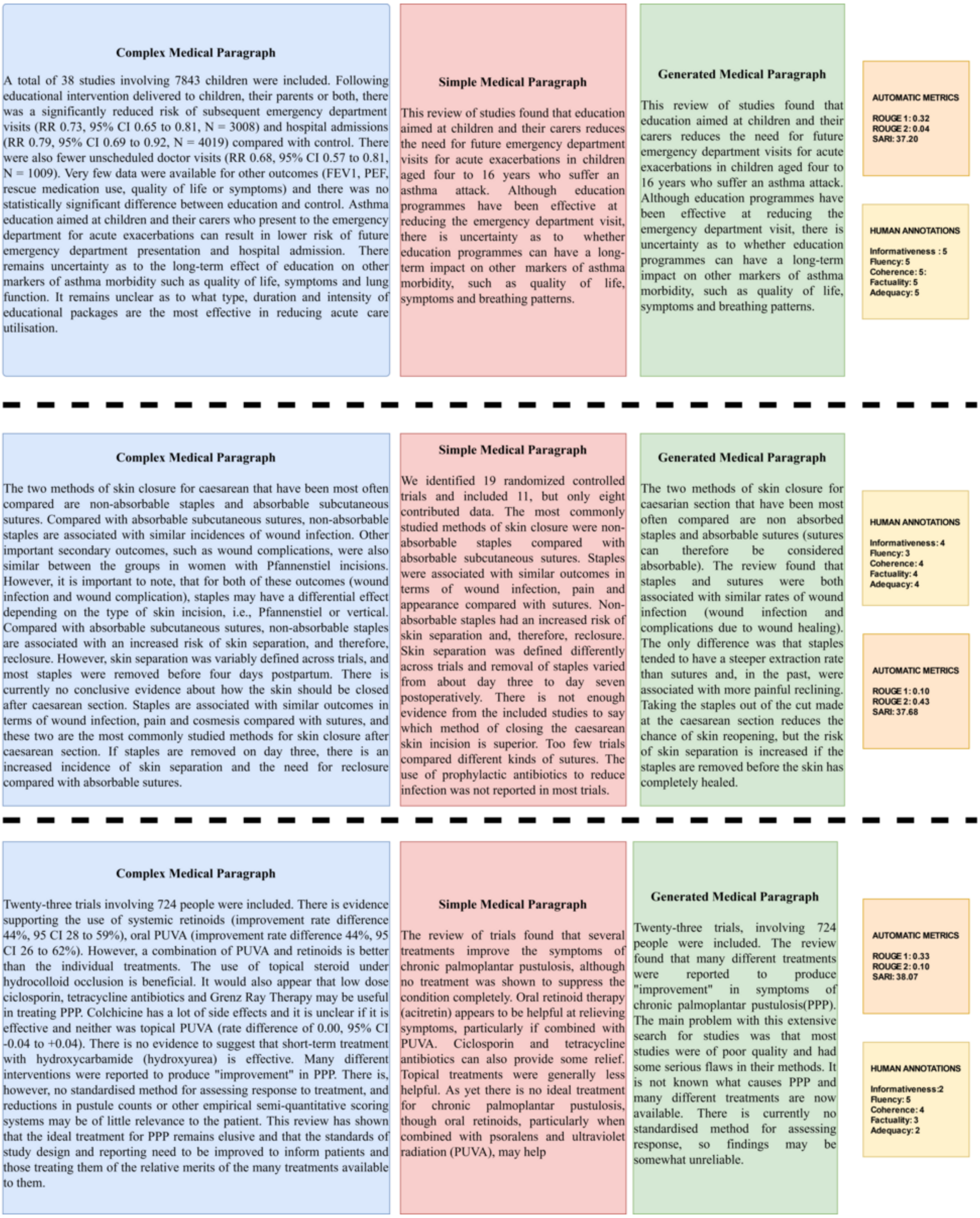
Rater	Informativeness	Fluency	Factuality	Coherence	Adequacy
A1	3.82	4.12	3.91	3.97	3.76
A2	3.50	4.97	3.59	4.82	3.68
A3	4.06	3.94	3.85	3.94	3.85
Average Likert score	3.79	4.34	3.78	4.24	3.76

Table 7. Spearman rank correlation coefficient between automatic metrics and human ratings for the text generated by TESLEA.

Metric	Informativeness	Fluency	Factuality	Coherence	Adequacy
ROUGE-1	0.18[a]	–0.04	–0.01	–0.05	0.06
ROUGE-2	0.08	–0.01	–0.05	–0.04	0.05
SARI	0.09	–0.66	–0.13	–0.01	0.01
BARTScore	0.08	0.32[a]	0.38[a]	0.22[a]	0.07[a]

[a] Best result.

Figure 7. Samples of Complex, Simple (Gold) and generated medical paragraphs along with automated metrics and Human annotations.



Complex Medical Paragraph

A total of 38 studies involving 7843 children were included. Following educational intervention delivered to children, their parents or both, there was a significantly reduced risk of subsequent emergency department visits (RR 0.73, 95% CI 0.65 to 0.81, N = 3008) and hospital admissions (RR 0.79, 95% CI 0.69 to 0.92, N = 4019) compared with control. There were also fewer unscheduled doctor visits (RR 0.68, 95% CI 0.57 to 0.81, N = 1009). Very few data were available for other outcomes (FEV1, PEF, rescue medication use, quality of life or symptoms) and there was no statistically significant difference between education and control. Asthma education aimed at children and their carers who present to the emergency department for acute exacerbations can result in lower risk of future emergency department presentation and hospital admission. There remains uncertainty as to the long-term effect of education on other markers of asthma morbidity such as quality of life, symptoms and lung function. It remains unclear as to what type, duration and intensity of educational packages are the most effective in reducing acute care utilisation.

Simple Medical Paragraph

This review of studies found that education aimed at children and their carers reduces the need for future emergency department visits for acute exacerbations in children aged four to 16 years who suffer an asthma attack. Although education programmes have been effective at reducing the emergency department visit, there is uncertainty as to whether education programmes can have a long-term impact on other markers of asthma morbidity, such as quality of life, symptoms and breathing patterns.

Generated Medical Paragraph

This review of studies found that education aimed at children and their carers reduces the need for future emergency department visits for acute exacerbations in children aged four to 16 years who suffer an asthma attack. Although education programmes have been effective at reducing the emergency department visit, there is uncertainty as to whether education programmes can have a long-term impact on other markers of asthma morbidity, such as quality of life, symptoms and breathing patterns.

AUTOMATIC METRICS

ROUGE 1: 0.32
ROUGE 2: 0.04
SARI: 37.20

HUMAN ANNOTATIONS

Informativeness : 5
Fluency: 5
Coherence: 5:
Factuality: 5
Adequacy: 5

Complex Medical Paragraph

The two methods of skin closure for caesarean that have been most often compared are non-absorbable staples and absorbable subcutaneous sutures. Compared with absorbable subcutaneous sutures, non-absorbable staples are associated with similar incidences of wound infection. Other important secondary outcomes, such as wound complications, were also similar between the groups in women with Pfannenstiel incisions. However, it is important to note, that for both of these outcomes (wound infection and wound complication), staples may have a differential effect depending on the type of skin incision, i.e., Pfannenstiel or vertical. Compared with absorbable subcutaneous sutures, non-absorbable staples are associated with an increased risk of skin separation, and therefore, reclosure. However, skin separation was variably defined across trials, and most staples were removed before four days postpartum. There is currently no conclusive evidence about how the skin should be closed after caesarean section. Staples are associated with similar outcomes in terms of wound infection, pain and cosmesis compared with sutures, and these two are the most commonly studied methods for skin closure after caesarean section. If staples are removed on day three, there is an increased incidence of skin separation and the need for reclosure compared with absorbable sutures.

Simple Medical Paragraph

We identified 19 randomized controlled trials and included 11, but only eight contributed data. The most commonly studied methods of skin closure were non-absorbable staples compared with absorbable subcutaneous sutures. Staples were associated with similar outcomes in terms of wound infection, pain and appearance compared with sutures. Non-absorbable staples had an increased risk of skin separation and, therefore, reclosure. Skin separation was defined differently across trials and removal of staples varied from about day three to day seven postoperatively. There is not enough evidence from the included studies to say which method of closing the caesarean skin incision is superior. Too few trials compared different kinds of sutures. The use of prophylactic antibiotics to reduce infection was not reported in most trials.

Generated Medical Paragraph

The two methods of skin closure for caesarian section that have been most often compared are non absorbed staples and absorbable sutures (sutures can therefore be considered absorbable). The review found that staples and sutures were both associated with similar rates of wound infection (wound infection and complications due to wound healing). The only difference was that staples tended to have a steeper extraction rate than sutures and, in the past, were associated with more painful reclining. Taking the staples out of the cut made at the caesarean section reduces the chance of skin reopening, but the risk of skin separation is increased if the staples are removed before the skin has completely healed.

HUMAN ANNOTATIONS

Informativeness: 4
Fluency: 3
Coherence: 4
Factuality: 4
Adequacy: 4

AUTOMATIC METRICS

ROUGE 1: 0.10
ROUGE 2: 0.43
SARI: 37.68

Complex Medical Paragraph

Twenty-three trials involving 724 people were included. There is evidence supporting the use of systemic retinoids (improvement rate difference 44%, 95 CI 28 to 59%), oral PUVA (improvement rate difference 44%, 95 CI 26 to 62%). However, a combination of PUVA and retinoids is better than the individual treatments. The use of topical steroid under hydrocolloid occlusion is beneficial. It would also appear that low dose ciclosporin, tetracycline antibiotics and Grenz Ray Therapy may be useful in treating PPP. Colchicine has a lot of side effects and it is unclear if it is effective and neither was topical PUVA (rate difference of 0.00, 95% CI -0.04 to +0.04). There is no evidence to suggest that short-term treatment with hydroxycarbamide (hydroxyurea) is effective. Many different interventions were reported to produce "improvement" in PPP. There is, however, no standardised method for assessing response to treatment, and reductions in pustule counts or other empirical semi-quantitative scoring systems may be of little relevance to the patient. This review has shown that the ideal treatment for PPP remains elusive and that the standards of study design and reporting need to be improved to inform patients and those treating them of the relative merits of the many treatments available to them.

Simple Medical Paragraph

The review of trials found that several treatments improve the symptoms of chronic palmoplantar pustulosis, although no treatment was shown to suppress the condition completely. Oral retinoid therapy (acitretin) appears to be helpful at relieving symptoms, particularly if combined with PUVA. Ciclosporin and tetracycline antibiotics can also provide some relief. Topical treatments were generally less helpful. As yet there is no ideal treatment for chronic palmoplantar pustulosis, though oral retinoids, particularly when combined with psoralens and ultraviolet radiation (PUVA), may help

Generated Medical Paragraph

Twenty-three trials, involving 724 people were included. The review found that many different treatments were reported to produce "improvement" in symptoms of chronic palmoplantar pustulosis(PPP). The main problem with this extensive search for studies was that most studies were of poor quality and had some serious flaws in their methods. It is not known what causes PPP and many different treatments are now available. There is currently no standardised method for assessing response, so findings may be somewhat unreliable.

AUTOMATIC METRICS

ROUGE 1: 0.33
ROUGE 2: 0.10
SARI: 38.07

HUMAN ANNOTATIONS

Informativeness:2
Fluency: 5
Coherence: 4
Factuality: 3
Adequacy: 2

Discussion

Principal Findings

The most up-to-date research about biomedicine is often inaccessible to the general public due to the domain-specific medical terminology. A way to address this problem is by creating a system that converts complex medical information into a simpler form, thus making it accessible to everyone. In this study, a TS approach was developed that can automatically simplify complex medical paragraphs while maintaining the quality of the generated text. The proposed approach trains the transformer-based BART model to optimize rewards specific for TS, resulting in increased simplicity. The BART model is trained using the proposed RL method to optimize certain rewards that help generate simpler text while maintaining the quality of generated text. As a result, the trained model generates simplified text that reduces the complexity of the original text by 2-grade points, when measured using the FKGL [29]. From the results obtained, it can be concluded that TESLEA is effective in generating simpler text compared with technical abstracts, the gold-standard references (ie, simple medical paragraphs corresponding to complex medical paragraphs), and the baseline models. Although previous work [8] developed baseline models for this task, to the best of our knowledge, this is the first time RL is being applied to the field of medical TS. Moreover, previous studies failed to analyze the quality of the generated text, which this study measures via the factors of FLU, FAC, COH, ADE, and INFO. Manual evaluations of TESLEA-generated text were conducted with the help of domain experts using the aforesaid factors and further research was conducted to analyze which automatic metrics agree with manual annotations using the Spearman rank correlation coefficient. The analysis revealed that BARTScore [45] best correlates with the human annotations when evaluated for a text generated by TESLEA, indicating that TESLEA learns to generate semantically relevant and fluent text, which conveys the essential information mentioned in the complex medical paragraph. These results suggest that (1) TESLEA can perform TS of medical paragraphs such that outputs are simple and maintain the quality, (2) the rewards optimized by TESLEA help the model capture syntactic and semantic information, increasing the FLU and COH of outputs, as witnessed when the outputs are evaluated by BARTScore and human annotators.

Limitations and Future Work

Although this research is a significant contribution to the literature on medical TS, the proposed approach does have a few limitations, addressing which can result in even better outputs. TESLEA can generate simpler versions of the text, but in some instances, it induces misinformation, resulting in reduced FAC and INFO of the generated text. Therefore, there is a need to design rewards that consider the FAC and INFO of the generated text. We also plan to conduct extensive human evaluations on a large scale for the text generated by various models (eg, KIS, BART-UL) using domain experts (ie, physicians and medical students).

Transformer-based language models are sensitive to the pretraining regime, so a possible next step is to pretrain a language model on domain-specific raw data sets such as PubMed [40], which will help develop domain-specific vocabulary for the model. Including these strategies may help in increasing the simplicity of the generated text.

Conclusion

The interest in and need for TS in the medical domain are of growing interest as the quantity of data is continuously increasing. Automated systems, such as the one proposed in this paper, can dramatically increase accessibility to information for the general public. This work not only provides a technical solution for automated TS, but also lays out and addresses the challenges of evaluating the outputs of such systems, which can be highly subjective. It is the authors' sincere hope that this work allows other researchers to build on and improve the quality of similar effort.

Acknowledgments

The authors thank the research team at DaTALab, Lakehead University, for their support. The authors also thank Compute Canada for providing the computational resources without which this research would not have been possible. This research is funded by NSERC Discovery (RGPIN-2017-05377) held by Dr. Vijay Mago. The authors thank Mr. Aditya Singhal (MSc student at Lakehead University) for providing his feedback on the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Training Procedures and Decoding Methods.
[DOCX File , 129 KB - medinform_v10i11e38095_app1.docx]

Multimedia Appendix 2
Hyperparameters and Evaluation Metrics.
[DOCX File , 190 KB - medinform_v10i11e38095_app2.docx]

Multimedia Appendix 3
Abbreviations and Examples.
[DOCX File , 1060 KB - medinform_v10i11e38095_app3.docx]

References

1. Carroll J, Minnen G, Pearce D, Canning Y, Devlin S, Tait J. Simplifying text for language-impaired readers. New Brunswick, NJ: Association for Computational Linguistics; 1999 Presented at: Ninth Conference of the European Chapter of the

Association for Computational Linguistics; June 8-12, 1999; Bergen, Norway p. 269-270 URL: https://aclanthology.org/E[]99-1042

2. Paetzold G, Specia L. Unsupervised Lexical Simplification for Non-Native Speakers. AAAI 2016 Mar 05;30(1):3761-3767 [FREE Full text] [doi: 10.1609/aaai.v30i1.9885]
3. Ganitkevitch J, Van Durme B, Callison-Burch C. PPDB: The paraphrase database. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Brunswick, NJ: Association for Computational Linguistics; 2013 Jun Presented at: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 9-12, 2013; Atlanta, GA p. 758-764 URL: https://aclanthology.org/N13-1092 [doi: 10.3115/v1/p15-2070]
4. Rebecca Thomas S, Anderson S. WordNet-Based Lexical Simplification of a Document. In: Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012). 2012 Presented at: The 11th Conference on Natural Language Processing (KONVENS 2012); September 19-21, 2012; Vienna, Austria p. 80 URL: https://www.researchgate.net/publication/270450791_WordNet-Based_Lexical_Simplification_of_a_Document
5. Qiang J, Li Y, Zhu Y, Yuan Y, Wu X. Lexical Simplification with Pretrained Encoders. AAAI 2020 Apr 03;34(05):8649-8656. [doi: 10.1609/aaai.v34i05.6389]
6. Zhu Z, Bernhard D, Gurevych I. A monolingual tree-based translation model for sentence simplification. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010 Organizing Committee; 2010 Presented at: The 23rd International Conference on Computational Linguistics (Coling 2010); August 23-27, 2010; Beijing, China p. 1353-1361 URL: https://aclanthology.org/C10-1152.pdf
7. Wubben S, van den Bosch A, Krahmer E. Sentence simplification by monolingual machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). New Brunswick, NJ: Association for Computational Linguistics; 2012 Presented at: The 50th Annual Meeting of the Association for Computational Linguistics; July 8-14, 2012; Jeju Island, Korea p. 1015-1024 URL: https://aclanthology.org/P12-1107
8. Devaraj A, Marshall I, Wallace B, Li J. Paragraph-level Simplification of Medical Texts. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Brunswick, NJ: Association for Computational Linguistics; 2021 Jun Presented at: The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6-11, 2021; Virtual p. 4972-4984 URL: https://aclanthology.org/2021.naacl-main.395.pdf [doi: 10.18653/v1/2021.naacl-main.395]
9. Nisioi S, Štajner S, Paolo Ponzetto S, Dinu LP. Exploring neural text simplification models. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). New Brunswick, NJ: Association for Computational Linguistics; 2017 Presented at: The 55th Annual Meeting of the Association for Computational Linguistics; July 30-August 4, 2017; Vancouver, BC p. 85-91 URL: https://aclanthology.org/P17-2014.pdf [doi: 10.18653/v1/p17-2014]
10. Afzal M, Alam F, Malik KM, Malik GM. Clinical Context-Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation. J Med Internet Res 2020 Oct 23;22(10):e19810 [FREE Full text] [doi: 10.2196/19810] [Medline: 33095174]
11. Zhang X, Lapata M. Sentence Simplification with Deep Reinforcement Learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. New Brunswick, NJ: Association for Computational Linguistics; 2017 Presented at: The 2017 Conference on Empirical Methods in Natural Language Processing; September 7-11, 2017; Copenhagen, Denmark p. 584-594 URL: https://aclanthology.org/D17-1062.pdf [doi: 10.18653/v1/d17-1062]
12. Xu W, Napoles C, Pavlick E, Chen Q, Callison-Burch C. Optimizing Statistical Machine Translation for Text Simplification. TACL 2016 Dec;4:401-415. [doi: 10.1162/tacl_a_00107]
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc; 2017 Presented at: NIPS'17: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 6000-6010 URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
14. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2020 Jul Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual p. 7871-7880 URL: https://aclanthology.org/2020.acl-main.703.pdf [doi: 10.18653/v1/2020.acl-main.703]
15. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Amazon AWS. 2022. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2022-10-31]
16. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 2020;21:1-67 [FREE Full text]
17. Martin L, Fan A, de la Clergerie E, Bordes A, Sagot B. MUSS: multilingual unsupervised sentence simplification by mining paraphrases. arXiv Preprint posted online on April 16, 2021. [doi: 10.48550/arXiv.2005.00352]

18. Wenzek G, Lachaux MA, Conneau A, Chaudhary V, Guzmán F, Joulin A, et al. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In: Proceedings of the Twelfth Language Resources and Evaluation Conference.: European Language Resources Association; 2020 Presented at: LREC 2020: The 12th Conference on Language Resources and Evaluation; May 11-16, 2020; Marseille, France p. 4003-4012 URL: https://aclanthology.org/2020.lrec-1.494
19. Zhao Y, Chen L, Chen Z, Yu K. Semi-Supervised Text Simplification with Back-Translation and Asymmetric Denoising Autoencoders. AAAI 2020 Apr 03;34(05):9668-9675. [doi: 10.1609/aaai.v34i05.6515]
20. Surya S, Mishra A, Laha A, Jain P, Sankaranarayanan K. Unsupervised Neural Text Simplification. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2019 Presented at: The 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy p. 2058-2068 URL: https://aclanthology.org/P19-1198.pdf [doi: 10.18653/v1/p19-1198]
21. Sun R, Jin H, Wan X. Document-Level Text Simplification: Dataset, Criteria and Baseline. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. New Brunswick, NJ: Association for Computational Linguistics; 2021 Presented at: The 2021 Conference on Empirical Methods in Natural Language Processing; November 7–11, 2021; Online and Punta Cana, Dominican Republic p. 7997-8013 URL: https://aclanthology.org/2021.emnlp-main.630.pdf [doi: 10.18653/v1/2021.emnlp-main.630]
22. Coster W, Kauchak D. Simple English Wikipedia: a new text simplification task. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. New Brunswick, NJ: Association for Computational Linguistics; 2011 Presented at: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; June 19-24, 2011; Portland, OR p. 665-669 URL: https://aclanthology.org/P11-2117.pdf
23. Jiang C, Maddela M, Lan W, Zhong Y. Neural CRF Model for Sentence Alignment in Text Simplification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2020 Jul Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Virtual p. 7943-7960 URL: https://aclanthology.org/2020.acl-main.709.pdf [doi: 10.18653/v1/2020.acl-main.709]
24. Xu W, Callison-Burch C, Napoles C. Problems in Current Text Simplification Research: New Data Can Help. TACL 2015 Dec;3:283-297. [doi: 10.1162/tacl_a_00139]
25. Bjerva J, Bos J, van der Goot R, Nissim M. The Meaning Factory: Formal Semantics for Recognizing Textual Entailment and Determining Semantic Similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). New Brunswick, NJ: Association for Computational Linguistics; 2014 Presented at: The 8th International Workshop on Semantic Evaluation (SemEval 2014); August 23-24, 2014; Dublin, Ireland p. 642-646 URL: https://aclanthology.org/S14-2114.pdf [doi: 10.3115/v1/s14-2114]
26. Laban P, Schnabel T, Bennett P, Hearst M. Keep It Simple: Unsupervised Simplification of Multi-Paragraph Text. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). New Brunswick, NJ: Association for Computational Linguistics; 2021 Presented at: The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; August 1–6, 2021; Online p. 6365-6378 URL: https://aclanthology.org/2021.acl-long.498.pdf [doi: 10.18653/v1/2021.acl-long.498]
27. van den Bercken L, Sips RJ, Lofi C. Evaluating neural text simplification in the medical domain. New York, NY: Association for Computing Machinery (ACM); 2019 May Presented at: WWW '19: The World Wide Web Conference; May 13-17, 2019; San Francisco CA p. 3286-3292 URL: https://dl.acm.org/doi/10.1145/3308558.3313630 [doi: 10.1145/3308558.3313630]
28. Dataset. Github. URL: https://github.com/AshOlogn/Paragraph-level-Simplification-of-Medical-Texts [accessed 2022-10-31]
29. Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch. 1975 Feb 1. URL: https://apps.dtic.mil/sti/citations/ADA006655 [accessed 2022-10-31]
30. Papineni K, Roukos S, Ward T, Zhu W. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. New Brunswick, NJ: Association for Computational Linguistics; 2002 Presented at: The 40th Annual Meeting of the Association for Computational Linguistics; July 7-12, 2002; Philadelphia, PA p. 311-318 URL: https://aclanthology.org/P02-1040.pdf [doi: 10.3115/1073083.1073135]
31. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. New York, NY: IEEE; 2019 Presented at: 2019 IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China p. 1-15 URL: https://ieeexplore.ieee.org/document/8904728 [doi: 10.1109/ICHI.2019.8904728]
32. Breland H. Word Frequency and Word Difficulty: A Comparison of Counts in Four Corpora. Psychol Sci 2016 May 06;7(2):96-99 [FREE Full text] [doi: 10.1111/j.1467-9280.1996.tb00336.x]
33. Narayan S, Cohen SB, Lapata M. Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. New Brunswick, NJ: Association for Computational Linguistics; 2018 Presented at: The 2018 Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium p. 1797-1807 URL: https://aclanthology.org/D18-1206.pdf [doi: 10.18653/v1/d18-1206]

34. Nallapati R, Zhou B, dos Santos C, Gu˙lçehre C, Xiang B. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. New Brunswick, NJ: Association for Computational Linguistics; 2016 Aug Presented at: The 20th SIGNLL Conference on Computational Natural Language Learning; August 7-12, 2016; Berlin, Germany p. 280-290 URL: https://aclanthology.org/K16-1028.pdf [doi: 10.18653/v1/k16-1028]
35. Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, et al. ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pretraining. In: Findings of the Association for Computational Linguistics, EMNLP 2020. New Brunswick, NJ: Association for Computational Linguistics; 2020 Presented at: EMNLP 2020; November 16-20, 2020; Online p. 2401-2410 URL: https://aclanthology.org/2020.findings-emnlp.217.pdf
36. Ranzato M, Chopra S, Auli M, Zaremba W. Sequence Level Training with Recurrent Neural Networks. arXiv Preprint posted online on May 6, 2016. [FREE Full text]
37. Aghajanyan A, Shrivastava A, Gupta A, Goyal N, Zettlemoyer L, Gupta S. Better Fine-Tuning by Reducing Representational Collapse. 2020 Apr Presented at: International Conference on Learning Representations (ICLR 2020); April 26–30, 2020; Virtual URL: https://www.researchgate.net/publication/343547031_Better_Fine-Tuning_by_Reducing_Representational_Collapse
38. Williams R. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 1992 May;8(3-4):229-256 [FREE Full text] [doi: 10.1007/BF00992696]
39. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-Critical Sequence Training for Image Captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY: IEEE; 2017 Jul Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI p. 7008-7024 URL: https://openaccess.thecvf.com/content_cvpr_2017/papers/Rennie_Self-Critical_Sequence_Training_CVPR_2017_paper.pdf [doi: 10.1186/isrctn12348322]
40. 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]
41. Martin L, De, Sagot B, Bordes A. Controllable Sentence Simplification. In: InProceedings of the 12th Language Resources and Evaluation Conference. 2020 May 11 Presented at: In Proceedings of the Twelfth Language Resources and Evaluation Conference; 2020-05-11; France p. 4689-4698 URL: https://aclanthology.org/2020.lrec-1.577/
42. Yan YY, Hu F, Chen J, Bhendawade N, Ye T, Gong Y, et al. FastSeq: Make Sequence Generation Faster. 2021 Aug 01 Presented at: InProceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Aug 2021; 2022-08-01; Thailand p. 218-226 URL: https://aclanthology.org/2021.acl-demo.26/ [doi: 10.18653/v1/2021.acl-demo.26]
43. Paulus R, Xiong C, Socher R. A Deep Reinforced Model for Abstractive Summarization. 2018 Presented at: International Conference on Learning Representations (ICLR 2018); April 30 to May 3, 2018; Vancouver, BC URL: https://www.researchgate.net/publication/316875315_A_Deep_Reinforced_Model_for_Abstractive_Summarization
44. Lin CY. ROUGE: A Package for Automatic Evaluation of Summarie. New Brunswick, NJ: Association for Computational Linguistics; 2004 Presented at: Text Summarization Branches Out; July 25 and 6, 2004; Barcelona, Spain p. 74-81 URL: https://aclanthology.org/W04-1013.pdf
45. Yuan W, Neubig G, Liu P. BARTScore: Evaluating Generated Text as Text Generation. 2021 May 21 Presented at: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021; December 6-14, 2021; Virtual p. 27263-27277 URL: https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html
46. Zhang J, Zhao Y, Saleh M, Liu P. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. 2020 Jul 13 Presented at: InInternational Conference on Machine Learning. 2020; 2020-07-13; Virtual URL: http://proceedings.mlr.press/v119/zhang20ae
47. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2018 Sep 27 Presented at: International Conference on Learning Representations; 2018; Vancouver, Canada.

Abbreviations

ARI: Automated Readability Index
BERT: bidirectional encoder representations from transformers
FKGL: Flesch-Kincaid Grade Level
GPT: generative pretraining transformer
MLE: maximum likelihood estimation
KIS: Keep it Simple
Lml: maximum likelihood loss
LS: lexical simplification
LSTM: long short-term memory
MUSS: multilingual unsupervised sentence simplification

PLS: plain language summary
RFlesch: FKGL reward
RL: reinforcement learning

Edited by T Hao; submitted 18.03.22; peer-reviewed by T Zhang, S Kim, H Suominen; comments to author 27.06.22; revised version received 08.08.22; accepted 12.10.22; published 18.11.22.

Please cite as:
Phatak A, Savage DW, Ohle R, Smith J, Mago V
Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning–Based Text Simplification Approach
JMIR Med Inform 2022;10(11):e38095
URL: https://medinform.jmir.org/2022/11/e38095
doi:10.2196/38095
PMID:36399375

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org