JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 9 (2021), Issue 10 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Review

Original Papers

Corrigenda and Addenda

Editorial

Review

Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study

Zheming Zuo[1], PhD; Matthew Watson[1], BSc; David Budgen[1], PhD; Robert Hall[2], BSc; Chris Kennelly[2], BSc, MBA; Noura Al Moubayed[1], PhD

[1]Department of Computer Science, Durham University, Durham, United Kingdom

[2]Cievert Ltd, Newcastle upon Tyne, United Kingdom

Corresponding Author:
Noura Al Moubayed, PhD
Department of Computer Science
Durham University
Lower Mountjoy, South Rd
Durham, DH1 3LE
United Kingdom
Phone: 44 1913341749
Email: Noura.al-moubayed@durham.ac.uk

Abstract

Background: Data science offers an unparalleled opportunity to identify new insights into many aspects of human life with recent advances in health care. Using data science in digital health raises significant challenges regarding data privacy, transparency, and trustworthiness. Recent regulations enforce the need for a clear legal basis for collecting, processing, and sharing data, for example, the European Union's General Data Protection Regulation (2016) and the United Kingdom's Data Protection Act (2018). For health care providers, legal use of the electronic health record (EHR) is permitted only in clinical care cases. Any other use of the data requires thoughtful considerations of the legal context and direct patient consent. Identifiable personal and sensitive information must be sufficiently anonymized. Raw data are commonly anonymized to be used for research purposes, with risk assessment for reidentification and utility. Although health care organizations have internal policies defined for information governance, there is a significant lack of practical tools and intuitive guidance about the use of data for research and modeling. Off-the-shelf data anonymization tools are developed frequently, but privacy-related functionalities are often incomparable with regard to use in different problem domains. In addition, tools to support measuring the risk of the anonymized data with regard to reidentification against the usefulness of the data exist, but there are question marks over their efficacy.

Objective: In this systematic literature mapping study, we aim to alleviate the aforementioned issues by reviewing the landscape of data anonymization for digital health care.

Methods: We used Google Scholar, Web of Science, Elsevier Scopus, and PubMed to retrieve academic studies published in English up to June 2020. Noteworthy gray literature was also used to initialize the search. We focused on review questions covering 5 bottom-up aspects: basic anonymization operations, privacy models, reidentification risk and usability metrics, off-the-shelf anonymization tools, and the lawful basis for EHR data anonymization.

Results: We identified 239 eligible studies, of which 60 were chosen for general background information; 16 were selected for 7 basic anonymization operations; 104 covered 72 conventional and machine learning–based privacy models; four and 19 papers included seven and 15 metrics, respectively, for measuring the reidentification risk and degree of usability; and 36 explored 20 data anonymization software tools. In addition, we also evaluated the practical feasibility of performing anonymization on EHR data with reference to their usability in medical decision-making. Furthermore, we summarized the lawful basis for delivering guidance on practical EHR data anonymization.

Conclusions: This systematic literature mapping study indicates that anonymization of EHR data is theoretically achievable; yet, it requires more research efforts in practical implementations to balance privacy preservation and usability to ensure more reliable health care applications.

(JMIR Med Inform 2021;9(10):e29871) doi:10.2196/29871

KEYWORDS

healthcare; privacy-preserving; GDPR; DPA 2018; EHR; SLM; data science; anonymization; reidentification risk; usability

Introduction

Background

Digital health [1] encompasses several distinct domains, including but not limited to automatic visual diagnostic systems [2], medical image segmentation [3], continuous patient monitoring [4], clinical data–driven decision support systems [5-7], connected biometric sensors [8,9], and expert-knowledge–based consultations [10,11] using personal electronic health records (EHRs) [12-14]. Of late, pervasive health care has become the central topic, attracting intensive attention and interest from academia [2-4], industry [5,10,11], and the general health care sector [13-15]. Developments achieved in the industry [5] and the health care sector [12-14,16] reveal the huge potential of data science in health care because of the common availability of medical patient data for secondary use (secondary use, also dubbed as reuse, of health care data refers to the use of data for a different purpose than the one for which the data were originally collected). However, such potential could be hindered by legitimate concerns over privacy [17].

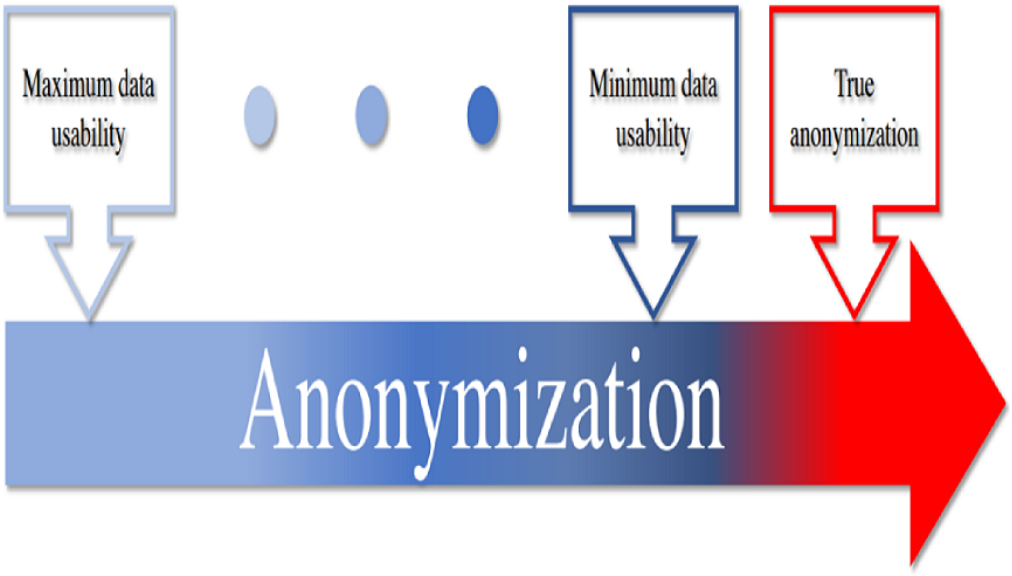
The United Kingdom's Human Rights Act 1998 defines privacy as "everyone has the right to respect for [their] private and family life, [their] home and [their] correspondence" in Article 8 [18]. However, it is difficult to explicitly define true privacy because of the discrepancies among target problems, for example, human-action recognition from videos [19], camera-pose estimation from images [20], and next-word prediction from articles [21]. In general, privacy can be treated as any personally identifiable information [22,23]. In the context of digital health care, the secondary use of patients' clinical data requires both the data controller (responsible for determining the purpose for which, and the means by which, health care data are processed) and data processor (responsible for processing health care data on behalf of the data controller) to comply with the lawful basis and gain direct consent from the data owner [24]. Recently, privacy invasion became an increasing concern in digital health care [25-28]. In 2014, the UK charity Samaritans (ie, data processor) released the app Radar [29] to identify potential distress and suicidality using the words and phrases of approximately 2 million Twitter (ie, data controller) users (ie, data owners). This app raised severe concerns among Twitter users, including those with a history of mental health issues, and thus it was pulled within weeks [26]. In 2015, the Royal Free London National Health Service (NHS) Foundation Trust (ie, data controller) shared 1.6 million complete and identifiable medical records of patients (ie, data owners) with DeepMind Technologies (Alphabet Inc; ie, data processor) to support further testing of the app Stream in assisting the detection of acute kidney injury [30]. This collaboration came under fire [27] for the inappropriate sharing of confidential patient data [24,31] and failure to comply with the United Kingdom's Data Protection Act (DPA), as was ruled [32] by the Information Commissioner's Office (ICO), which cited missing patient consent as well as lack of detailed purpose of use, research ethics approval, and the necessary process transparency [25]. Thus, a prerequisite for secondary use of clinical patient data is to guarantee patient privacy through data anonymization [33]. This is supported by legislation established in different countries that states that secondary use of clinical patient data is permitted if, and only if, the exchanged information is sufficiently anonymized in advance to prevent any possible future association with the data owners (ie, patients) [28,34]. For instance, researchers from academia pointed out the importance of patient-specific health data, which became the impetus for updating the United States' Health Information Portability and Accountability Act (HIPAA) in 2003 [35,36].

On January 30, 2020, a declaration [37] by the World Health Organization named the COVID-19 [38] outbreak a Public Health Emergency of International Concern. At present (as of April 18, 2021), there are a total of 140,835,884 and 4,385,938 confirmed cases and 3,013,111 and 150,419 deaths, respectively, throughout the world [39] and the United Kingdom [40]. As COVID-19 spread to every inhabitable continent within weeks [41], data science research relating to digital health care through large-scale data collection [42,43] and crowdsourcing [44,45] has been highly recommended to curb the ongoing pandemic, including virus tracing [46,47] and contact tracing [48,49]. Public concern with respect to privacy has significantly increased amid the COVID-19 pandemic [50,51]. For instance, mobile apps have been adopted to make contact tracing and notification instantaneous upon case confirmation [52,53], for example, the latest NHS COVID-19 app [54]. This is typically achieved by storing a temporary record of proximity events among individuals and thus immediately alerting users of recent close contact with diagnosed cases and prompting them to self-isolate. These apps have been placed under public scrutiny over issues of data protection and privacy [48].

Currently, the lack of more intuitive guidance and a deeper understanding of how to feasibly anonymize personally identifiable information in EHRs (it should be noted that data from wearables, smart home sensors, pictures, videos, and audio files, as well as the combination of EHR and social media data, are out of the scope of this study) while ensuring an acceptable approach for both patients and the public leave the data controller and data processor susceptible to breaches of privacy. Although several diligent survey papers [55-58] have been published to ensure privacy protection and suppress disclosure risk in data anonymization, sensitive information still cannot be thoroughly anonymized by reducing the risk of reidentification while still retaining the usefulness of the anonymized data—*the curse of anonymization* (Figure 1). Concretely, the gaps in the existing survey studies are four-fold: (1) there does not exist a single data anonymization survey that considers lawful aspects such as the European Union's General Data Protection Regulation (GDPR) as well as the DPA, ICO, and health care provider regulations; (2) most existing survey studies do not focus on digital health care; (3) the existing privacy models are usually incomparable (particularly for the values of parameters) and have been proposed for different

problem domains; and (4) the most recent trends of privacy model–based and machine learning–based data anonymization tools have not been summarized with adequate discussions in terms of their advantages and disadvantages. Motivated by these observations, we aim to deliver a clear picture of the landscape of lawful data anonymization while mitigating its curse in pervasive health care.

Figure 1. The curse of anonymization. Blue hue indicates an increase in data anonymity, which, in turn, reveals the decrease in usability of the anonymized data, very likely reaching minimum usability before reaching full anonymization (red hue).



A Brief Overview of the Problem Domain

Private Data and Their Categorization

In line with the updated scope of the GDPR and its associated Article 9 [59,60], private (ie, personal) data are defined as any direct or indirect information related to an identified or identifiable natural person. In general, based on the definition and categorization presented in chapter 10 of *Guide to the De-Identification of Personal Health Information* by El Emam [61], there are 5 types of data: relational data, transactional data, sequential data, trajectory data, and graph data. In addition, inspired by the survey study by Zigomitros et al [62], we also included image data because an EHR is essentially a 2D data matrix and thus could be viewed as a 2D image and anonymized using statistical and computer vision techniques.

Relational data [62] are the most common type of data. This category usually contains a fixed number of variables (ie, columns) and data records (ie, rows). Each data record usually pertains to a single patient, with that patient appearing only once in the data set. Typical relational data in health care can include clinical data in a disease or population registry. Transactional data [63] have a variable number of columns for each record. For instance, a data set of follow-up appointments from a hospital may consist of a set of prescription drugs that were prescribed to patients, and different patients may have a different number of transactions (ie, appointments) and prescribed drugs in each transaction. Sequential data [64] are similar to transactional data, but there is an order to the items in each record. For instance, a data set containing *Brachytherapy planning time* would be considered sequential data because some items appear before others. Sequential data can also be termed relational-transactional data. Trajectory data [65] combine sequential data with location information. For instance, data on the movement of patients would have location and timestamp information. Trajectory data can also be termed geolocational data. Graph data [66] encapsulate the relationships among objects using techniques from graph theory. For instance, data showing telephone calling, emailing, or instant messaging patterns between patients and general practitioners (GPs) could be represented as a graph, with patients and GPs being represented as nodes and a call between a given patient and their GP represented as an edge between their respective nodes. Graph data are also commonly used in social media [67]. Image data, as tabular medical records (ie, EHRs), can be treated as a grayscale image in 2D space. It should be noted that, in this study, the term image data does not refer to medical images such as computed tomography scans.

Types of Identifiers

How the attributes are handled during the anonymization process depends on their categorization [61]. All attributes contained in a table X are usually grouped into 4 types: direct identifying attributes I, indirect identifying attributes (ie, quasi-identifiers [QIs]) Q, sensitive attributes S, and other attributes O [61]. Refer to Multimedia Appendix 1 for the mathematical symbols and definitions used throughout this study.

Direct identifiers I, which are also termed direct identifying attributes, provide explicit links to data subjects and can be used to directly identify patients [68]. In practice, one or more direct

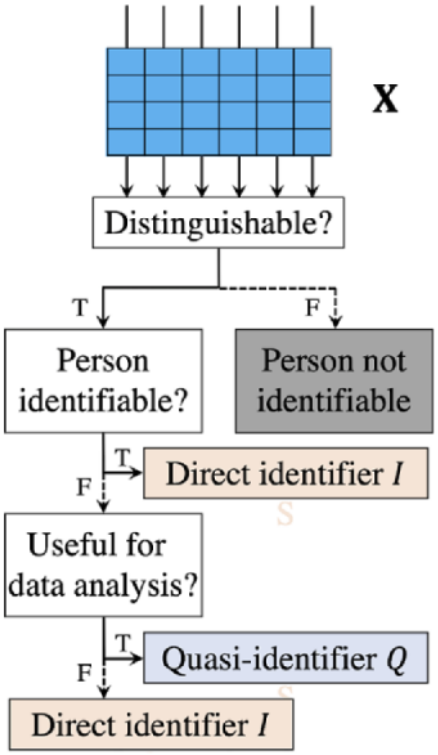
identifying attributes can be assigned to uniquely identify a patient, either by themselves or in conjunction with other information sources. Typical examples of the former case include NHS number, national insurance number, biometric residence permit number, and email address. Suppose there are 2 patients with the same full name within a single NHS foundation trust, the attribute *full name* cannot be a direct identifier by itself. However, a combination of *full name* and *living address* will be a direct identifier.

Indirect identifiers *Q*, or QIs, are identifiers that, when used with background knowledge of patients in the anonymized data set, can be used to reidentify a patient record with a high probability. Note that if someone, say, an adversary, does not have background knowledge of patients at hand, then this attribute cannot be deemed a QI. In addition, a common choice of QI also considers the analytical utility of the attribute. That is, a QI is usually useful for data analysis, whereas a direct identifier is not [61]. Typical QIs include gender, date of birth, postcode, and ethnic origin.

Sensitive attributes *S* are not useful with respect to the determination of the patient's identity; yet, they contain sensitive health-related information about patients, such as clinical drug dosage. Other attributes *O* represent variables that are not considered sensitive and would be difficult for an adversary to use for reidentification.

Among the 4 categories of identifiers, it is particularly difficult to differentiate between direct identifiers *I* and QIs *Q* In general, there are 3 determination rules used for this purpose [61], which are depicted in Figure 2: (1) an attribute can be either *I* or *Q* if it can be known by an adversary as background knowledge; (2) an attribute must be treated as *Q* if it is useful for data analysis and as *I* otherwise; and (3) an attribute should be specified as *I* if it can uniquely identify an individual.

Figure 2. Logical flow of distinguishing direct identifiers I from quasi-identifiers Q. F: false; T: true.



In Multimedia Appendix 2 [69,70], we summarize the features that are commonly listed as direct and indirect identifiers by health care bodies [71] that guide, inform, and legislate medical data release. All listed features may lead to personal information disclosure, and the list is by no means exhaustive. As more varied health care data are released and explored, more identifiers will be added to the lists of those featured in common data attack strategies, such as those in the studies by Hrynaszkiewicz et al [69] and Tucker et al [70], 18 HIPAA identifiers [72], and policies published by the NHS [73] and its foundation trusts, for example, Kernow [74] and Solent [75].

Data Anonymization Versus Data Pseudonymization

Given the definition in Recital 26 [76] of the most recent GDPR update, data anonymization (the term is common in Europe, whereas deidentification is more commonly used in North America) is a useful tool for sharing personal data while preserving privacy. Anonymization can be achieved by changing identifiers through removal, substitution, distortion, generalization, or aggregation. In contrast, data pseudonymization is a data management and deidentification procedure by which personally identifiable information fields within a data record are replaced by one or more artificial identifiers or pseudonyms.

It should be noted therefore that the relationship between data anonymization and pseudonymization techniques is characterized as follows:

- Anonymized data are not identifiable, whereas pseudonymized data are identifiable.
- Pseudonymized data remain personal based on Recital 26 of the GDPR and the conclusion [77] provided by the ICO.
- Solving the problem of data anonymization necessarily means solving pseudonymization.

Concretely, given an anonymization function *A* and raw data X, we have the anonymized data X'=*A*(X) such that there does not exist another function *R* that reidentifies the raw data X from the anonymized data X', that is, *R*(X')=*R*(*A*(X))=X. If such a function does exist, this is pseudonymization. The difference between these 2 operations can be generalized as follows: X→X' for anonymization and X→X'.

In a real-world scenario, efficient data anonymization is challenging because it is usually problem dependent (ie, solutions vary across problem domains) and requires substantial domain expertise (eg, to specify the direct and indirect identifiers present in raw data) and effort (eg, user involvement in specifying the privacy model before the data anonymization process). Fundamentally, it is very challenging and nontrivial to define what *true anonymization* is or, equivalently, to determine whether the raw data have been adequately anonymized (as well as to agree upon the definition of *adequate anonymization*). In practice, as visualized in Figure 1, we observe that as the level of data anonymity increases, the usability of the anonymized data decreases and very likely reaches minimum usability before reaching full anonymization. This fact combined with the need for more accurate models in health care provides sufficient motivation for continued research into methods of data anonymization. For this study, we believe that how anonymization is defined is problem dependent. We reiterate that there is no clear-cut line between pseudonymization and anonymization because even anonymized data can practically have different reidentification risks [78,79] (depending on the type of anonymization performed).

Aims of the Study

Objectives

To minimize bias and deliver up-to-date studies related to data anonymization for health care, we organized this survey in a systematic literature mapping (SLM) manner. In general, there are 2 main approaches to conduct literature reviews: systematic literature review (SLR) and SLM [80-82]. SLRs aim to identify, classify, and evaluate results to respond to a specific review question (RQ), whereas SLMs seek to investigate multiple RQs. In addition, SLRs synthesize evidence and consider the strength of such evidence [83], whereas an SLM provides an overview of a research area by reviewing the topics that have been covered in the literature [84]. Concretely, we combined high-quality systematic review studies—provided in the Cochrane Database of Systematic Reviews [85], Manchester; Centre for Reviews and Dissemination [86], York; and Health Technology Assessment [87], National Institute for Health Research—to explain this work explicitly and concisely with respect to the validity, applicability, and implication of the results.

Our overall objective is to alleviate the issues introduced toward the end of the previous section by reviewing the landscape of data anonymization for digital health care to benefit practitioners aiming to achieve appropriate trade-offs in leveraging the reidentification risk and usability of anonymized health care data. In other words, we evaluate the evidence regarding the effectiveness and practicality of data anonymization operations, models, and tools in secondary care from the perspective of data processors.

Defining RQs

The aims of the study are to evaluate the potential of preserving privacy using data anonymization techniques in secondary care. Concretely, we, as data processors, are highly motivated to investigate the best possible way of anonymizing real-world EHRs by leveraging the privacy and usability concerns visualized in Figure 1. Therefore, our RQs were defined as follows:

- RQ 1: Do best practices exist for the anonymization of realistic EHR data?
- RQ 2: What are the most frequently applied data anonymization operations, and how can these operations be applied?
- RQ 3: What are the existing conventional and machine learning–based privacy models for measuring the level of anonymity? Are they practically useful in handling real-world health care data? Are there any new trends?
- RQ 4: What metrics could be adopted to measure the reidentification risk and usability of the anonymized data?
- RQ 5: What are the off-the-shelf data anonymization tools based on conventional privacy models and machine learning?

The knowledge generated from this SLM, especially the answer to our driving question, RQ 1, will build on the study's evidence on the future of the development of data anonymization toolkits for data processors such as the companies and organizations in which they are situated. The evidence gained may also contribute to our understanding of how data anonymization tools are implemented and their applicability to anonymizing real-world health care data. Finally, we intend to identify the major facilitators and barriers to data anonymization in secondary care in relation to reidentification risk and utility.

Methods

Research Design

Data Sources and Search Strategy

In keeping with our RQs, we built up our search strategy using *keywords and indexing terms* and *Boolean operators*; the former refers to the general terms used when searching, and the latter represents the restrictions on these terms. Example keywords and indexing terms used included domain-specific terms such as *healthcare*, *digital health*, *digital healthcare*, *health monitoring*, and *eHealth*; problem-specific terms such as *data anonymization*, *anonymizer*, *de-identification*,

privacy-preserving, and *data protection*; data-specific terms such as *electronic medical records*, *electronic health records (EHR)*, *DICOM/CT images*, and *videos*; disease-specific terms such as *brain tumor*, *cervical cancer*, *breast cancer*, and *diabetes*; organization-specific terms such as *NHS*, *ICO*, *NIHR*, and *MRC*; and law-specific terms such as *DPA*, *GDPR*, and *HIPAA*. Example Boolean operators are *AND* and *OR*. Next, to avoid bias and ensure reliability, 2 researchers (ZZ and MW) used Google Scholar, Web of Science, Elsevier Scopus, and PubMed for searching academic studies up to June 2020; these services were used because they encompass a wide spectrum of databases such as IEEE Xplore, SpringerLink, ACM Digital Library, Elsevier Science Direct, arXiv, *The BMJ*, *Lancet*, and the *New England Journal of Medicine*. In addition, to maximize search coverage, we conducted forward and backward *snowball sampling* [88] (snowball sampling refers to using the reference list of a selected paper [backward snowballing] or the citations of a selected paper [forward snowballing]) on the selected studies. In particular, because gray literature is an important source of SLRs and SLMs [89] and they play a primary role in health care [90,91], gray literature was used to initialize our search in this study. Concretely, preprints from non–peer-reviewed electronic archives (eg, arXiv) or early-stage research were examined and distinguished in the follow-up study selection phase.

Inclusion and Exclusion Criteria

Articles were eligible for inclusion based on the criteria defined in Textbox 1. ZZ and MW assessed articles independently for inclusion eligibility. Inclusion is relatively straightforward in comparison with exclusion which can be more sweeping. Therefore, further clarification regarding some of the exclusion criteria is required. For instance, *without Experiment section* denotes that the article does not report on any evaluation of the ideas it contains using real-word clinical data sets. *Insights not suitable for EU/UK* indicates observations delivered by articles that treat personally identifiable data as a commercial commodity, as is the practice in, for example, the United States [92]. Preprints (tier 2 gray literature [93]) were carefully considered for selection in line with the inclusion and exclusion criteria summarized in Textbox 1. For duplicate articles (eg, a conference article that extended to a journal paper or a preprint paper accepted by either a conference or a journal), including those with a different title but essentially the same content, we only retained the publication with the highest quality to avoid double counting. To this end, we preferred to retain the article published by the journal with the highest impact factor. In the worst case, none of the duplicates would have been selected if they were all conference papers because this would have been a breach of research ethics.

Textbox 1. Inclusion and exclusion criteria for article selection.

Inclusion criteria

- Related to anonymization or privacy-preserving techniques
- Related to privacy-preserving techniques in health care
- Presented privacy concerns in health care
- Proposed methods for privacy preservation in electronic health records
- Proposed methods for using private information, for example, biometric data
- Proposed methods partially related to protected health care

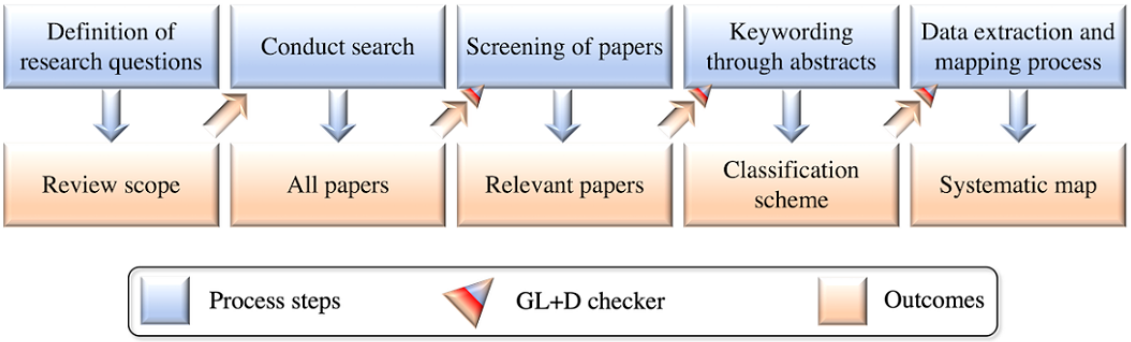
Exclusion criteria

- Written in language other than English
- Without *Abstract* or *Experiment* section
- About other health care issues, for example, clinical trials
- Insights not suitable for European Union or United Kingdom
- Out of our research scope
- Duplicate articles (case dependent)

Article Selection Phases

Article selection (Figure 3) consisted of 5 phases: (1) initially, we searched Google Scholar, Web of Science, Elsevier Scopus, and PubMed; (2) next, we applied the inclusion-exclusion criteria to the returned results from the initial search, including the qualifying preprints; (3) we then read the included articles and removed the irrelevant articles; (4) next, we conducted forward and backward snowball sampling on highly related articles; (5) finally, we double-checked the excluded articles and added relevant ones. In addition, we used the *GL+D Checker* mechanism shown in Figure 3, which refers to a combination of a *Gray Literature Checker* and a *Duplicates Checker*, each of which could also be used separately, depending on the situation.

Figure 3. Systematic literature mapping process for articles. GL+D: gray literature and duplicates.



Data Anonymization Toolkit Selection Phases

As mentioned at the beginning of this section, the phases involved in selecting data anonymization software tools are difficult because of the limited tools available in the existing studies. Thus, the initially included tools were selected from the qualified articles without considering whether their source code was publicly accessible, maintainable, and extensible. The only criterion was whether the tool could be downloaded and executed. Furthermore, to guarantee that the selection process was less biased, we decided that in each of the 2 (ie, privacy model–based and machine learning–based) categories of privacy-preserving software tools, the number of tools chosen from outside of the selected articles would be no more than 30% of the total.

Conduct of the Study

Qualified Articles

In keeping with the five-phase article selection strategy described in the previous section, ZZ and MW independently selected articles for eligibility in phase 2. Articles were moved forward to the *Article reading* phase or excluded after a full agreement was reached. In addition, NAM served as an arbitrator for any unresolved disagreement. The selection process was conducted using 3 consecutive steps: (1) the title and abstract of each article were screened for relevance; (2) full article contents were reviewed for those without certainty for inclusion; and (3) forward and backward snowballing was applied to the remaining articles to maximize search coverage. The full reference list of the included articles and the related systematic review or mapping studies were also screened by hand for additional articles. There were a total of 13 preprints among the 192 selected articles (Figure 4) after phase 1. Before beginning phase 2, by applying the *Gray Literature Checker* mechanism, we observed that 4 of the 13 preprints had been successfully published in either peer-reviewed conferences [94-96] or journals [97]. Next, the *Duplicates Checker* was applied consecutively to remove their preprint versions. Using the same process in each phase, we accumulated a total of 239 articles to include in this SLM study, including 9 preprints. Details of the 239 selected research articles are grouped in categorical order and chronological order in Table 1 and Figure 5, respectively.

Figure 4. Number of selected articles during the study selection process.

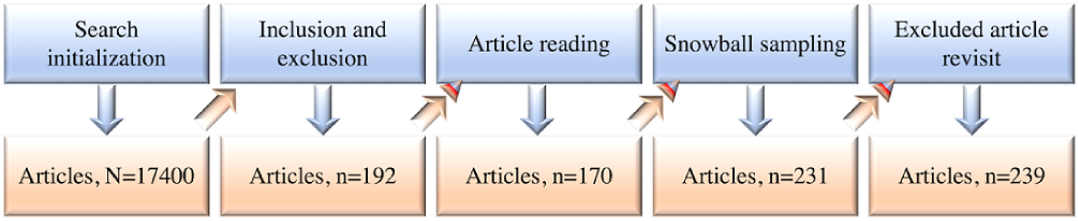
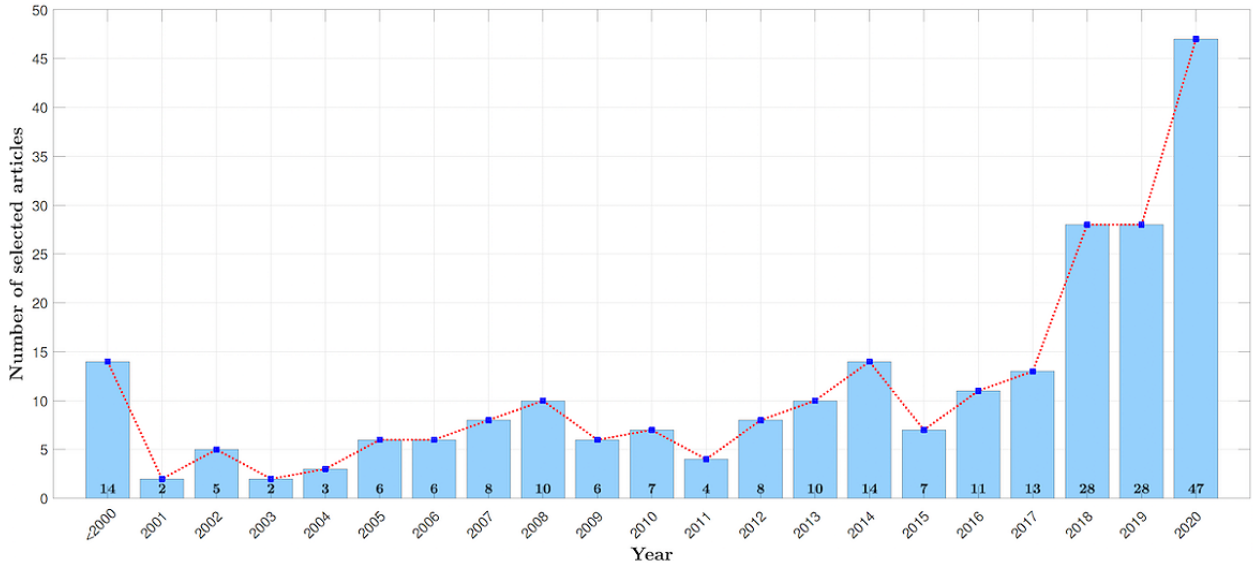


Table 1. An overview of the 239 selected research articles grouped in categorical order.

Category	Selected research articles, n (%)
Background knowledge	60 (25.1)
Data anonymization operations	16 (6.7)
Privacy models	104 (43.5)
Risk metrics	4 (1.7)
Utility metrics	19 (7.9)
Data anonymization tools	36 (15.1)

Figure 5. An overview of the 239 selected research articles grouped in chronological order.



Qualified Software Tools

In accordance with the strategy of selecting qualified privacy-preserving software tools described in the previous section, there were 5 out of a total of 15 privacy model–based data anonymization tools that were not derived from the qualified (ie, selected) articles. Of these 5 tools, 3 (*Amnesia*, OpenAIRE; *Anonimatron*, realrolfje; and *Anonymizer*, Divante Ltd) were obtained by searching GitHub [98], and the remaining 2 (*OpenPseudonymiser*, Julia Hippisley-Cox and *NLM-Scrubber* from the US National Library of Medicine) were found through Google Search. Of the 5 machine learning–based tools, only one (*CrypTen*, Facebook Inc) was obtained from GitHub.

Results

Four Categories

To add structure to this SLM, we grouped the results of the reviewed articles into four categories: *Basic Data Anonymization Operations* (for RQ 2), *Level of Anonymity Guarantees and Evaluations* (for RQ 3), *Disclosure Risk Assessments* and *Usability Measurements* (for RQ 4), and *Existing Privacy Model–Based Data Anonymization Tools*, *Existing Machine Learning–Based Data Anonymization Tools*, and *Legal Framework Support* (for RQ 5). RQ 1, as the leading RQ, is answered in *Results Summary for RQs*.

Basic Data Anonymization Operations

Perturbation

This technique is implemented by modifying the original data in a nonstatistically significant fashion. As described in the code of practice [99] provided by the ICO, the alteration of values within the data set should decrease the vulnerability of that data set to data linkage. The benefit of this method is that it anonymizes the raw data while guaranteeing that the statistical usefulness of the data remains unchanged. On this basis, the possible drawback of such a method is the accuracy of the anonymized data.

This technique can be achieved through, for instance, microaggregation [100], data swapping [101] (equivalent to permutation [102]), rank swapping [103]), postrandomization [104], adding noise [105], and resampling [106], all of which are described, with real-world health care examples to explain each operation, in Multimedia Appendix 3 [100,101,104-109]. For microaggregation, an observed value is replaced with the average value calculated over a small group of units. The units belonging to the same group are represented by the same value in the anonymized data. This operation can be applied independently to a single variable or to a set of variables with the original column or columns removed. For data swapping, the data records are altered through the switching of variable values across pairs of records in a fraction of the raw data. Equivalently, permutation rearranges the values (either randomly or systematically) and is useful where mapping to alternate configurations of alphanumeric values is problematic or redundant. To this end, the raw data can be efficiently perturbed by permuting the sensitive attribute and the value of a similar record. This operation not only guarantees the statistical significance of the anonymized data but also reduces the risk of the record-wise reidentification. For postrandomization, categorical variables are perturbed based on a prescribed probability mechanism such as a Markov matrix. For raw numerical data with low variance, adding noise, that is, adding a random value, is commonly adopted. Alternatively, resampling is also frequently used on raw numerical data by drawing repeated samples from the original data.

Generalization

Generalization [107] relies on an observable attribute having an underlying hierarchy. This is an example of such a typical hierarchy:

Full postcode → street → city or town → county (optional) → country

with a possible instance being as follows:

DH1 3LE → South Road → Durham → UK

and

DH → Durham → UK

Typically, generalization is used to reduce the specificity of the data and thereby the probability of information disclosure. Given the examples above, the degree of generalization is fully controlled by the granularity defined in the hierarchy.

Suppression

Suppression [110] refers to local suppression in data anonymization research. This is usually achieved by replacing the observed value of one or more variables with *missing* or *NA* or -. This method helps to address problems where rows would be dropped because of the difficulty of successfully applying perturbation or other generalization methods to guarantee their inclusion in the anonymized data set. By suppressing categorical values that render the rows identifiable, useful data from those rows will not be lost. This method can only be used when the raw data are varied enough that they prevent the suppressed value from being inferred.

Data Masking

Data masking [108] is a technique frequently used for creating a structurally similar yet inauthentic version of the raw data. This technique helps to protect the original sensitive data while providing a functional substitute and should be used in settings in which the original raw data are not required.

Differential Privacy

Differential privacy (DP) [109] aims to help organizations better understand the requirements of end users by maximizing the accuracy of search queries while minimizing the probability of identifying personal data information. This is achieved in practice by performing techniques such as data filtering, adaptive sampling, adding noise by fuzzifying certain features, and analyzing or blocking intrusive queries. Essentially, a DP algorithm updates values, leaving some intact while replacing others such that a potential attacker is unable to determine whether a value is fake or genuine. For details about practical DP and related techniques, please refer to section 1.4 of Multimedia Appendix 4 [57,66,111-165].

Homomorphic Encryption

Homomorphic encryption (HE) [166] is a technique that enables calculations to be performed on encrypted data directly, without the need to decrypt the data. The drawbacks of such a method are slow execution speeds. To the best of our knowledge, and in accordance with the definitions used in this paper, a technique that uses an encryption method cannot be treated as anonymization. The presence of the *key* makes the data theoretically reversible and therefore constitutes data pseudonymization. A well-known extension of HE is termed additive HE, which supports secure addition of numbers given only the encrypted data [167].

Compressive Privacy

Compressive privacy (CP) [168] is a technique that proposes to perform privatization by mapping the original data into space with a lower dimension. This is usually achieved by extracting the key features required for the machine learning model before sending the data to the cloud server. To this end, data owners (eg, NHS trusts and authorized companies) have control over privacy [169]. Alternatively, this technique could be performed before applying the chosen privacy models. Essentially, CP can be treated as a dimensionality reduction technique that also preserves privacy. Privacy models related to CP are presented in the following section.

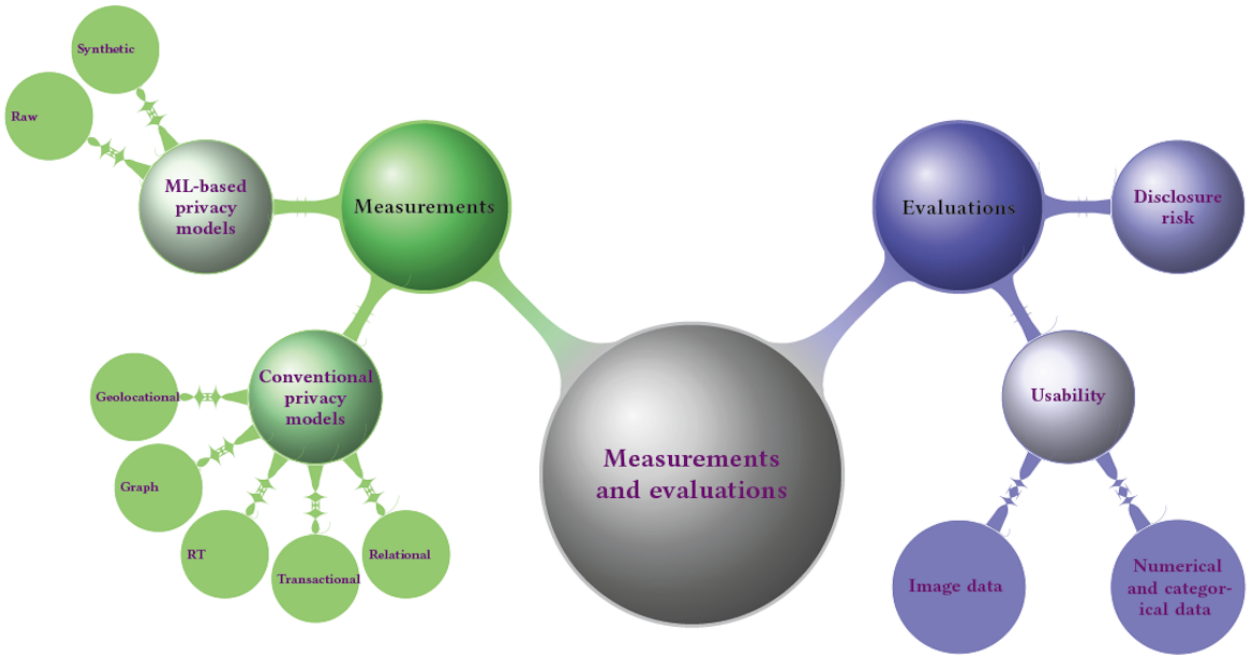
Level of Anonymity Guarantees and Evaluations

Measurement and Evaluation

Two Models

The objective of satisfying different levels of anonymity is usually achieved through 2 consecutive steps: measurement and evaluation. The former refers to the use of either conventional or machine learning–based privacy models to perform data anonymization, and the latter is the process of evaluating the reidentification risk and degree of usability of the anonymized data. The anonymization operations are usually adopted by conventional privacy models or machine-learning–based models. Figure 6 provides a way to quickly locate content of interest.

Figure 6. Categorizations of measurements and evaluations for achieving different levels of anonymity. ML: machine learning; RT: relational-transactional privacy model.



Conventional Privacy Models

The attributes contained in a table are usually divided into direct identifiers *I*, QIs *Q*, and sensitive identifiers *S*. The direct identifiers *I* are usually removed at the very beginning stage of data anonymization. Thus, a table X required to be anonymized is denoted as X(*S*, *Q*).

Given a class of records *G* in a table X, we want to create a single equivalent group *C* using a function *A* such that *C*=*A*(*G*) or *C'*=*A*(*C*). The monotonicity property of privacy models is defined for a single equivalent group *C* or class of records *G*. This property is required by several models for the purpose of refining the level of anonymization of *C*. This property is also useful for manipulating anonymized data by converting it into coarse-grained classes with equivalent classes (ie, a set of anonymized data records that share the same *Q*). This is a simple and computationally inexpensive solution. However, it would be inefficient, particularly in a case where the anonymized data are released to several organizations, each of which has a different minimum acceptable degree of anonymity. To this end, it is always a good practice to first perform the anonymization and then generate multiple coarser versions of the data, rather than performing separate anonymization for each organization [170].

During the process of data anonymization, interpretable and realistically feasible measurements (ie, privacy models [171]) should be considered to measure the level of anonymity of the anonymized data. The off-the-shelf privacy models (summarized as part of Figure 7) are usually independent of any data deanonymization attack and measure the privacy level using features of the anonymized data. One step further, 35 conventional privacy models were investigated to support data with the types grouped into 5 categories (Table 2).

Figure 7. General pipeline for existing privacy model–based data anonymization tools. F: false; T: true.

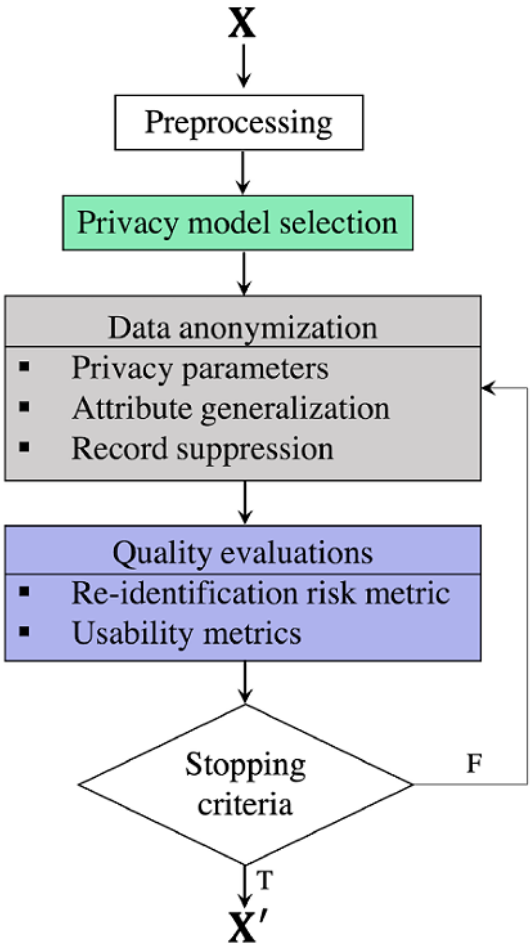


Table 2. A summary of privacy models for relational electronic health record data with respect to parameter interval and degree of privacy of data.

Category	Privacy model	Section in Multimedia Appendix 4	Parameter interval	Privacy level	References
Relational					
	κ-anonymity	1.1	[1, |X|]	High	[111-117,172,173]
	(α, *k*)-anonymity	1.1.1	$\alpha \in [0, 1]$, $k \in [0, +\infty]$	α: low, *k*: high	[114,174]
	k-map	1.1.2	[1, |X|]	Low	[112,175]
	m-invariance	1.1.3	$[0, +\infty]$	High	[176]
	(*k*, *e*)-anonymity	1.1.4	$[0, +\infty]$	High	[57,118,177]
	(*k*, *g*)-anonymity	1.1.5	$k \in [0, +\infty]$, $g \in [0, 1]$	High	[178,179]
	Multirelational *k*-anonymity	1.1.6	$[0, +\infty]$	High	[180]
	Strict average risk	1.1.7	N/A[a]	Low	[181,182]
	l-diversity	1.2	$[0, +\infty]$	High	[119]
	l^+-diversity	1.2.1	$l \in [0, +\infty]$, $\theta \in [0, 1]$	High	[120]
	t-closeness	1.3	$[0, +\infty]$	Low	[121,122]
	Stochastic *t*-closeness	1.3.1	$t \in [0, +\infty]$, $\varepsilon \in [0, +\infty]$	Low	[123]
	(*c*, *t*)-isolation	1.3.2	$[0, +\infty]$	High	[124]
	β-Likeness and enhanced β-likeness	1.3.3	$[0, +\infty]$	High	[125]
	Differential privacy	1.4	$[0, +\infty]$	Low	[109]
	(*k*, ε)-anonymity	1.4.1	$[0, +\infty]$	High	[126-131]
	(ε, δ)-anonymity	1.4.2	$\varepsilon \in [0, +\infty]$, $\delta \in [0, +\infty]$	ε: low, δ: low	[132-137]
	(ε, *m*)-anonymity	1.4.3	$\varepsilon \in [0, 1]$, $m \in [1, +\infty]$	ε: high, *m*: high	[118]
	Distributed differential privacy	1.4.4	$[0, +\infty]$	Low	[138]
	Distributional differential privacy	1.4.5	$\varepsilon \in [0, +\infty]$, $\delta \in [0, +\infty]$	ε: low, δ: low	[139]
	d-χ-privacy	1.4.6	$[0, +\infty]$	Low	[140]
	Joint differential privacy	1.4.7	$\varepsilon \in [0, +\infty]$, $\delta \in [0, +\infty]$	ε: low, δ: low	[183]
	(*X*, *Y*)-anonymity	1.5.1	[0, 1]	Low	[141]
	Normalized variance	1.5.2	[0, 1]	High	[142]
	δ-disclosure privacy	1.5.3	$[0, +\infty]$	High	[143]
	(*d*,*y*)-privacy	1.5.4	[0, 1]	Low	[144,145]
	δ-presence	1.5.5	[0, 1]	Low	[57,146]
	Population and sample Uniqueness	1.5.6 or 1.5.7	N/A	N/A	[79,147-151]
	Profitability	1.5.8	N/A	N/A	[152]
Transactional	k^m-anonymity	2	N/A	N/A	[153]
Relational-transactional	(*k*, k^m)-anonymity	3	N/A	N/A	[154]
Graph					
	k-degree	4.1	N/A	N/A	[155-158]
	k^2 degree	4.2	N/A	N/A	[156]
	k-automorphism	4.3	N/A	N/A	[157,159,160]
	(*k*, *l*)-anonymity	4.4	N/A	N/A	[66,161,162]

Category	Privacy model	Section in Multimedia Appendix 4	Parameter interval	Privacy level	References
Geolocational	Historical *k*-anonymity	5	N/A	N/A	[163]

[a]N/A: not applicable.

Machine Learning–Based Privacy Models

Two Categories

In light of machine learning and its derived subset, deep learning, there has been an upsurge of interest in machine learning– or deep learning–based privacy models for anonymizing patient or general data; we explore these approaches in this section. We divided related machine learning–based privacy models into 2 categories in accordance with the type of data used: raw or synthetic. Of late, the use of synthetic data has become more popular because these generated data are both anonymous and realistic; therefore, consent from data owners is not required [184]. The data in this category can be generated using techniques such as generative adversarial networks (GANs) [185] and usually do not have the risk of reidentification; thus, research works concentrate on improving the utility of synthetic data.

Models for Raw Data

In the study by D'Acquisto and Naldi [186], conventional principal component analysis (PCA) was used to anonymize sensitive data sets to achieve anonymization-utility trade-offs, that is, maximize both the information loss and utility. Different from its use in reducing the dimension of the data, where the smallest principal components are removed, PCA was instead adopted to remove the largest principal components before data projection. To measure the usefulness of the data anonymized through PCA, several utility metrics were presented; these are discussed in detail in Multimedia Appendix 5 [117,172,186-213]. In the domain of data anonymization, the first work using PCA is termed as differentially private PCA [214]. This technique explores the trade-off between the privacy and utility of low-rank data representations by guaranteeing DP. The study by Dwork et al [215] suggested that noise be added directly to the covariance matrix before projection in PCA.

Many similar PCA techniques rely on results derived from random matrix theory [216-219]. To reduce the computational cost of the privacy model, additive HE was used for PCA with a single data user [217], where the rank of PCA with an unknown distribution could be adaptively estimated to achieve (,)-DP [218]. More recently, the concept of collaborative learning (or shared machine learning) [94,97,220] became very popular in data anonymization. That is, the data collected from multiple parties are collectively used to improve the performance of model training while protecting individual data owners from any information disclosure. For instance, both HE and secret sharing were adopted in privacy-preserving PCA [219] for horizontally partitioned data, that is, data sets share the same feature space but different sample space. In that work, HE could be substituted with secure multiparty computation (SMPC) [221] for industrial use (more details are provided in *SMPC Frameworks* under *Results*).

Despite the great success achieved by PCA and its variants in data anonymization, traditional clustering algorithms have also been adopted to deal with the same problem; -means [222], fuzzy c-means [223,224], Gaussian mixture model [225,226], spectral clustering [227,228], affinity propagation [229], and density-based spatial clustering of applications with noise [230,231] are some of the algorithms that have been used for data anonymization. Most recently, anonymization solutions were proposed for privacy-preserving visual tasks in color images. For instance, the conventional -nearest neighbor algorithm was combined with DP [232] for privacy-preserving face attribute recognition and person reidentification. Homomorphic convolution was proposed by combining HE and secret sharing [233] for visual object detection, and adversarial perturbation was devised to prevent disclosure of biometric information in finger-selfie images [234].

Models for Synthetic Data

In the study by Choi et al [95], GANs were adopted to generate realistic synthetic patient records (medical GAN [medGAN]; [235]) by learning the distribution of real-world multilabel discrete EHRs. Concretely, medGAN was proposed to generate multilabel discrete patient records through the combination of an autoencoder and a GAN. Such a network supports the generation of both binary and numeric variables (ie, medical codes such as diagnosis, medication, and procedure codes) and the arrangement of records in a matrix format where each row corresponds to a patient and each column represents a specific medical code. The study by Baowaly et al [236] extended the original medGAN by using both Wasserstein GANs with gradient penalty [237] and boundary-seeking GANs [96] to speed up model convergence and stability. In addition, GANs have also been used for segmenting medical images (ie, brain magnetic resonance imaging scans) while coping with privacy protection and data set imbalances [238]. In other words, GANs have proven their potential in data augmentation for imbalanced data sets and data anonymization for privacy preservation. A conditional GAN framework— anonymization through data synthesis-GAN [239]—was proposed to generate synthetic data while minimizing *patient identifiability*, which is based on the probability of reidentification given the combination of all data of any individual patient. In addition, DP has also been used in conjunction with GANs to generate synthetic EHRs [240-243]; most of these models were summarized in a recent survey [244]. On the basis of the CP technique introduced in the previous section, the study by Tseng and Wu [245] presented compressive privacy generative adversarial network to provide a data-driven local privatization scheme for creating compressed representations with lower dimensions for cloud services while removing sensitive information from raw images. Most recently, the conditional identity anonymization GAN [246] was proposed for image and video anonymization based on conditional GANs [247]. Concretely, conditional identity anonymization GAN supports the removal of identifiable information such as

characteristics of human faces and bodies while guaranteeing the quality (granularity) of the generated images and videos.

Disclosure Risk Assessments

Given the conventional and machine learning–based privacy models, a disclosure risk assessment is usually conducted to measure the reidentification risk of the anonymized EHR data. In practice, risk values from different combinations of privacy models could be used when deciding which version of the anonymized data should be used for data analysis and possible machine learning tasks such as EHR classification with respect to treatment planning or distance recurrence identification.

Concretely, there are 3 major types of disclosure that may occur during the process of data anonymization: identity, attribute, and membership disclosure (Table 3). For practical guidance, we have provided a comparative summary in Multimedia Appendix 6 [248-251] of most of the 35 conventional privacy models investigated (in terms of parameter value ranges and privacy levels).

Table 3. Categorization of data reidentification risk metrics for electronic health record data.

Disclosure type and metric		Section in Multimedia Appendix 6	Reference
Identity			
	Average risk	1	N/A[a]
	Overall risk	1	N/A
	β-Likeness	1	[125]
	Distance-linked disclosure	2	[248]
Attribute			
	Probabilistic linkage disclosure	2	[249]
	Interval disclosure	2	[250]
Membership	Log-linear models	3	[251]

[a]N/A: not applicable.

Usability Measurements

The metrics used for measuring the usefulness of the anonymized data can be treated as an on-demand component of a data anonymization system. We revisit the proposed quantitative metrics in this section, although this important indicator is usually not fully covered in the off-the-shelf privacy model–based data anonymization tools. In addition, qualitative metrics are not covered in this study. This is due to the varied objectives of different data anonymization activities, including the evaluation of anonymization quality that is performed by health care professionals. Table 4 lists the selected data usability metrics and the type of data for which they are suitable.

Table 4. Categorization of data usability metrics.

Data type and metric	Section in Multimedia Appendix 5	References
Numerical and categorical		
Information loss and its variants	1.1	[172,187-189]
Privacy gain	1.2	[190]
Discernibility	1.3	[191]
Average equivalence class size	1.4	[117]
Matrix norm	1.5	[192,193]
Correlation	1.6	[194]
Divergence	1.7	[195,196]
Image[a]		
Mean squared error and its variants	2.1	[197-200]
Peak signal-to-noise ratio	2.2	[201-206]
Structural similarity index	2.3	[207,208]

[a]Any type of raw and anonymized electronic health record data that can be converted into an image.

Existing Privacy Model–Based Data Anonymization Tools

In this section, several off-the-shelf data anonymization tools based on conventional privacy models and operations are detailed. These tools are commonly adopted for anonymizing tabular data. It should be noted that EHRs are usually organized in the tabular data format and that the real difficulties of anonymizing tabular data lie in the inherent bias and presumption of the availability of limited forecast-linkable data. Therefore, we investigated 14 data anonymization toolboxes, all of which share a similar workflow (summarized in Figure 8 and compared in Table 5 and Table 6). Functionally similar toolboxes are introduced together below.

Figure 8. Overall results of the systematic literature mapping study. This mapping consists of four contextually consecutive parts (from bottom to top): basic anonymization operations, existing privacy models, metrics proposed to measure re-identification risk and degree of usability of the anonymized data, and off-the-shelf data anonymization software tools. ADS-GAN: anonymization through data synthesis using generative adversarial networks; AP: affinity propagation; BL: β-Likeness; CIAGAN: conditional identity anonymization generative adversarial network; CPGAN: compressive privacy generative adversarial network; DBSCAN: density-based spatial clustering of apps with noise; DP: differential privacy; DPPCA: differentially private principal component analysis; FCM: fuzzy c-means; G: graph; GAN: generative adversarial network; GL: geolocational; GMM: Gaussian mixture model; HE: homomorphic encryption; IL: information loss; ILPG: ratio of information loss to privacy gain; KA: *k*-Anonymity; *k*NN+DP: *k*-nearest neighbor+differential privacy; LD: *l*-Diversity; medGAN: medical generative adversarial network; ML: machine learning; PCA: principal component analysis; PG: privacy gain; PPPCA: privacy-preserving principal component analysis; R: relational; RT: relational-transactional; SC: spectral clustering; T: transactional; TC: *t*-Closeness.

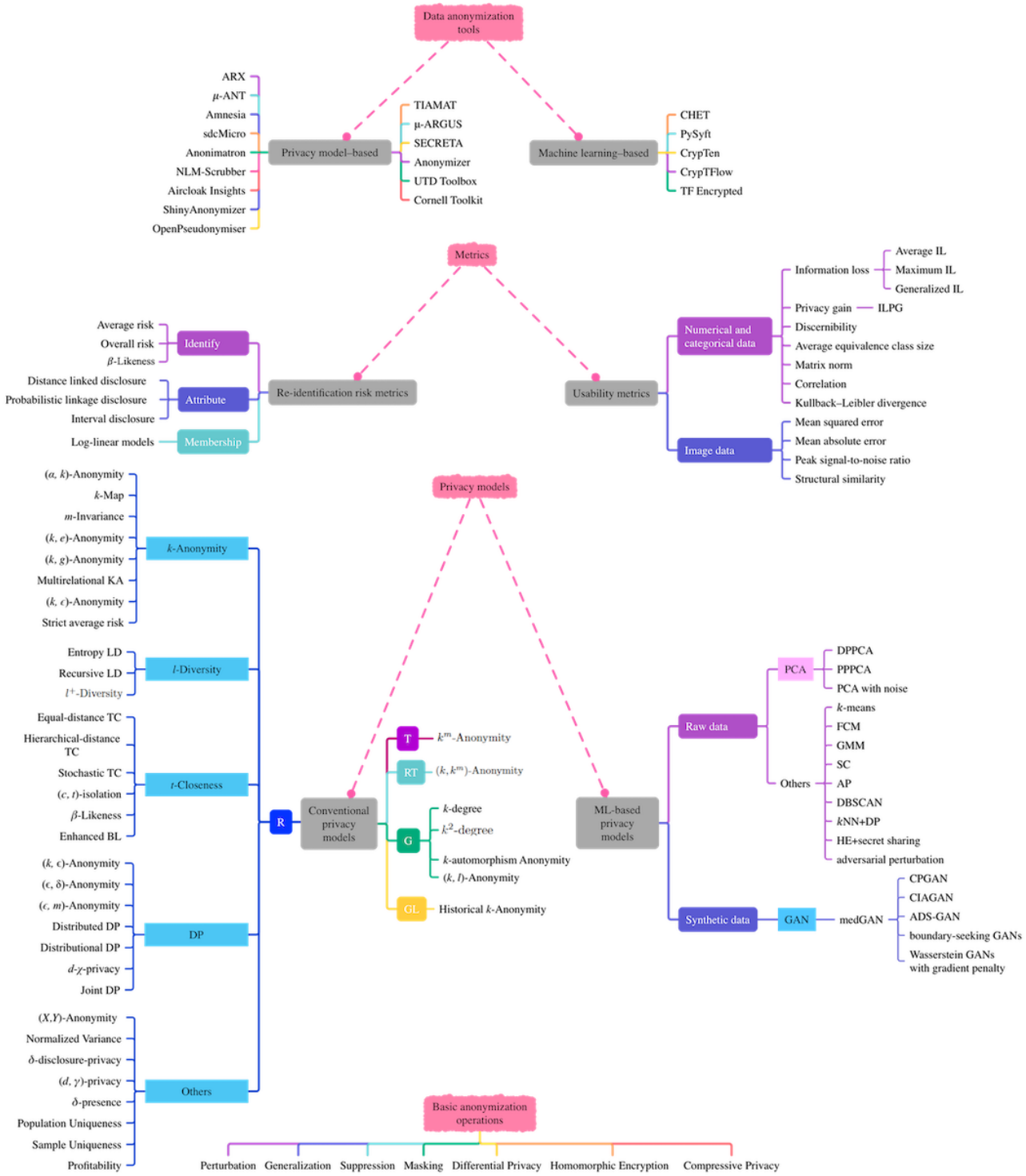


Table 5. Comparison of the off-the-shelf privacy model–based data anonymization tools in terms of available development options, anonymization functionality and risk metrics.

Tool	Last release	Development support					Anonymization	Risk assessment
		Open source	Public API[a]	Extensibility	Cross-platform	Programming language		
ARX	November 2019	✓[b]	✓	✓	✓	Java	✓	✓
Amnesia	October 2019	✓	✓	✓	✓	Java	✓	
μ-ANT[c]	August 2019	✓	✓	✓	✓	Java	✓	
Anonimatron	August 2019	✓	✓	✓	✓	Java		
SECRETA[d]	June 2019				✓	C++	✓	
sdcMicro	May 2019	✓	✓	Poorly supported	✓	R	✓	✓
Aircloak Insights	April 2019				✓	Ruby		
NLM[e] Scrubber	April 2019				✓	Perl		
Anonymizer	March 2019	✓	✓	✓	✓	Ruby		
Shiny Anonymizer	February 2019	✓	✓	✓	✓	R	✓	
μ-ARGUS	March 2018					C++	✓	✓
UTD[f] Toolbox	April 2010	✓		Poorly supported	✓	Java	✓	
OpenPseudonymiser	November 2011	✓			✓	Java		
TIAMAT[g]	2009				✓	Java	✓	
Cornell Toolkit	2009	✓		Poorly supported	✓	C++	✓	Poorly supported

[a]API: application programming interface.

[b]Feature present.

[c]μ-ANT: microaggregation-based anonymization tool.

[d]SECRETA: System for Evaluating and Comparing RElational and Transaction Anonymization.

[e]NLM: National Library of Medicine.

[f]UTD: University of Texas at Dallas.

[g]TIAMAT: Tool for Interactive Analysis of Microdata Anonymization Techniques.

Table 6. Comparison of the off-the-shelf privacy model–based data anonymization tools with respect to the supported privacy models.

Tool	Last release	Privacy models									
		k-anonymity	*l*-diversity	*t*-closeness	δ-presence	*k*-map	(*k*, *g*)-anonymity	(*k*, ε)-anonymity	(ε, δ)-anonymity	k^m-anonymity	(k, k^m)-anonymity
ARX	November 2019	✓[a]	✓	✓	✓	✓			✓		
Amnesia	October 2019	✓									
μ-ANT[b]	August 2019	✓		✓							
Anonimatron	August 2019										
SECRETA[c]	June 2019	✓								✓	✓
sdcMicro	May 2019	✓	✓								
Aircloak Insights	April 2019										
NLM[d] Scrubber	April 2019										
Anonymizer	March 2019										
Shiny Anonymizer	February 2019										
μ-ARGUS	March 2018	✓									
UTD[e] Toolbox	April 2010	✓	✓	✓							
OpenPseudonymiser	November 2011										
TIAMAT[f]	2009	✓	✓	✓							
Cornell Toolkit	2009		✓	✓							

[a]Feature present.

[b]μ-ANT: microaggregation-based anonymization tool.

[c]SECRETA: System for Evaluating and Comparing RElational and Transaction Anonymization.

[d]NLM: National Library of Medicine.

[e]UTD: University of Texas at Dallas.

[f]TIAMAT: Tool for Interactive Analysis of Microdata Anonymization Techniques.

Amnesia [252] supports 2 privacy models, k^m-anonymity and *k*-anonymity; the former is used for set-valued and relational-set data sets, and the latter is used for tabular data. Amnesia does not support any reidentification risk assessment; the authors claim that there is no risk associated with the anonymized data set because every query on the anonymized attributes will return at least *k* records.

Anonimatron [253] state that it has been GDPR-compliant since 2010. It supports working with several databases out of the box. It can also be used with text files. The software conducts search-and-replace tasks based on custom rules and as such is merely a pseudonymization tool; however, it is extensible because of its open-source nature.

ARX [164,181,254] was originally developed for biomedical data anonymization. In terms of conventional privacy models, ARX mainly supports 6 additional privacy models: (1) strict average risk, (2) population uniqueness, (3) sample uniqueness, (4) δ-disclosure privacy, (5) β-likeness, and (6) profitability. The population uniqueness can be measured using 4 different models described in the studies by Pitman [148], Zayataz [149], Chen and McNulty [150], and Dankar et al [255]. For *t*-closeness, there are 3 different variants for categorical and numeric data. The profitability privacy model is a game-theoretic model used to conduct cost-benefit analysis and maximize the monetary gains of the data publisher [152]. ARX is open source [256] and supports data of high dimensionality. It is available

as a library to be integrated into custom projects or as an installable graphical user interface tool. Similar to ARX, the microaggregation-based anonymization tool (μ-ANT) [257] is also open source [258] and extensible. μ-ANT supports 2 privacy models, k^m-anonymity and *k*-anonymity, as well as *t*-closeness [122]. With respect to usability measurements, μ-ANT supports both information loss and sum of the squared errors. However, μ-ANT does not support functions for filling the missing attribute values (this requires manual data preprocessing, instead, either by removal or filling with average values) or metrics to evaluate the reidentification risk of the anonymized data.

sdcMicro [259] supports 2 privacy models (*k*-anonymity and *l*-diversity) in conjunction with recoding, suppression, postrandomization method (PRAM; which works on categorical data and is usually treated as encompassing noise addition, data suppression, and data recoding. Specifically, each value of a categorical attribute is mapped to a different value in accordance with a prescribed Markov matrix, that is, PRAM matrix), noise addition, and microaggregation. Apart from these functions, this tool also supports the measurement of reidentification risk. As a tool similar to sdcMicro, μ-ARGUS [260] has been implemented in multiple programming languages. It supports anonymization of both microdata and tabular data. It is packaged as disclosure control software and includes *k*-anonymity, recoding (generalization), suppression, PRAM, noise, and microaggregation. Compared with sdcMicro and μ-ARGUS, both University of Texas at Dallas Toolbox [261] and Tool for Interactive Analysis of Microdata Anonymization Techniques [262] support 3 privacy models but lack a risk-assessment module. In addition, University of Texas at Dallas Toolbox was compared with ARX in the study by Prasser et al [263] because of their similar automated anonymization processes and perspectives (in both, the data set is treated as population data, describing one individual per record). In this comparison, ARX showed better performance with respect to execution times and measured data utility.

SECRETA (System for Evaluating and Comparing RElational and Transaction Anonymization) [264] handles 3 categories of data: relational data, transactional data, and relational-transactional data, which are respectively supported by *k*-anonymity and its variants, k^m-anonymity and (k, k^m)-anonymity. For relational data sets, SECRETA supports various schemes for data generalization, including full-domain generalization, subtree generalization, and multidimensional generalization. For transactional data, it supports k^m-anonymity using hierarchy-based generalization and constraint-based generalization. For measuring the risk of reidentification, the standalone Identification of Privacy Vulnerabilities toolkit [265] is used.

Aircloak Insights [266,267] can be deemed a data pseudonymization tool because it does not tackle any task of data anonymization. Concretely, by investigating 2 research studies [266,267], we argue that Aircloak Insights is focused more on data protection than on data anonymization. Aircloak Insights comes with a Diffix backend [267], which is essentially a middleware proxy to add noise to user queries for database access in an encrypted fashion. This is also inconsistent with what the authors announced on their official website: “Our privacy-preserving analytics solution uses patented and proven data anonymization that provides GDPR-compliant and high-fidelity insights” [268]. Nevertheless, a number of summarized attacks [267] may be used for validating the efficiency and efficacy of the data anonymization toolbox associated with the Aircloak pipeline.

National Library of Medicine-Scrubber [269] is an anonymization software tool that is specifically designed for coping with unstructured clinical text data. As such, *k*-anonymity is not applicable. Privacy is achieved by applying the HIPAA Safe Harbor model. National Library of Medicine-Scrubber treats text data anonymization as a process of eliminating a specific set of identifiers from the data, and the level of anonymization depends on the comprehensiveness of the identifier lookup data source. In addition, the reidentification risk measurement is not considered in this tool because the authors think that there is no established measure for reidentification of the patient from an anonymized text document.

OpenPseudonymiser [270] and ShinyAnonymizer [271] are very similar: both conduct data encryption only, although they have been specifically designed for medical data. As they only perform data encryption, they are not adequate for data anonymization. Concretely, they support a number of hashing functions (eg, MD5 and SHA512) and encryption algorithms (eg, data encryption standard and advanced encryption standard). Although they support several fundamental data anonymization operations (eg, removing information, suppression, generalization, and bottom and top coding), they do not implement any of the operations in line with privacy models. In addition, they do not provide tools for calculating the risk of reidentification or the measurement of data utility. Similarly, Anonymizer [272] was introduced as a universal tool to create anonymized databases. This tool replaces all data in the given database with anonymized random data where the unique, alphanumeric values are generated by the MD5 hashing function. To this end, the anonymized data might be less useful in view of the authors’ announcement [273]: “There is no way to keep nonanonymized rows in a table”; thus, this software tool is useful for database randomization rather than anonymization.

The Cornell Toolkit [274] supports *l*-diversity and *t*-closeness with flexible parameter configurations. Although the software supports the ability to display the disclosure risk of reidentification of the original tabular data, the method or methods used for implementing the risk measurement have not been introduced in either the paper [274] or in the documentation on the web [275], leaving this software with a low degree of explainability and, hence, trustworthiness.

Existing Machine Learning–Based Data Anonymization Tools

Two Classes

Recently, in response to the GDPR and DPA regulations, efforts were made by the machine learning and cryptography communities to develop privacy-preserving machine learning methods. We define privacy-preserving methods as any machine

learning method or tool that has been designed with data privacy as a fundamental concept (usually in the form of data encryption) and that can typically be divided into 2 classes: those that use SMPC and those that use fully HE (FHE). All the investigated machine learning–based data anonymization tools are compared in Table 7.

Table 7. Comparison of existing machine learning–based data anonymization tools. The Largest model tested column reports the number of parameters in the largest model shown in the respective tool's original paper (when reported); CrypTFlow has been shown to work efficiently on much larger machine learning models than the other available privacy-preserving machine learning tools.

Tool	Encryption methods				Reidentification risk assessment	Usability measurement	Development support		
	SMPC[a]	FHE[b]	Differential privacy	Federated learning			Supports training	Malicious security	Largest model tested
CrypTen	✓[c]						✓		N/A[d]
TF Encrypted	✓	✓		✓			✓		419,720
PySyft	✓		✓	✓			✓		N/A
CrypTFlow	✓							✓	65×10^6
CHET		✓							421,098

[a]SMPC: secure multiparty computation.

[b]FHE: fully homomorphic encryption.

[c]Feature present.

[d]N/A: not applicable.

SMPC Frameworks

SMPC involves a problem in which n parties, each with their own private input $x_1, x_2,..., x_n$ where party i has access to input x_i (and only x_i), wish to compute some function $f(x_1, x_2,..., x_n)$ without revealing any information about their private data [276] to the other parties. Most SMPC frameworks assume the parties to be semihonest: under this scheme we assume that malicious parties still follow the set protocol (although they may work together to attempt to extract private information). The current state-of-the-art framework for SMPC is SPDZ [277], and it is upon this framework that many SMPC-based machine learning libraries are built. This allows data owners to keep their data private and also allows for the machine learning model to be hidden. However, it does require at least three trusted, noncolluding parties or servers to work together to provide the highest level of protection; this can mean it is difficult to implement in practice. There are also significant overheads with this method; not only do SPDZ algorithms necessarily take longer to compute (because of cryptographic overhead), but there is also a significant amount of communication that needs to take place among all participating parties. This results in SMPC machine learning models running approximately 46 times slower than plaintext variants [278], meaning that it is impractical to use such models with large and complex data sets.

There are several different practical implementations of this type of protocol, although none are ready for use in production environments. CrypTen [279] is a library that supports privacy-preserving machine learning with PyTorch. CrypTen currently supports SMPC (although support for other methods such as FHE is in development) by providing SMPC-encrypted versions of tensors and many PyTorch functions; it also includes a tool for encrypting a pre-existing PyTorch model. Although CrypTen supports many of PyTorch's existing functions, it still has certain limitations. Most notably, it does not currently support graphics processing unit computation, which significantly hinders its ability to be used in conjunction with large, complex models. TensorFlow (TF) Encrypted [280] is a similar framework for the TF open-source software library for machine learning that also supports SMPC through the SPDZ framework. TF Encrypted also includes support for federated learning (which allows the training of machine learning models to be distributed over many devices without each device needing to reveal its private data) and HE.

PySyft [278] is a more general framework than CrypTen or TF Encrypted because it supports multiple machine learning libraries (including TF and PyTorch) and multiple privacy methods. As part of this, it features SMPC-based machine learning, much like CrypTen and TF Encrypted, but also allows for additional layers of security to be incorporated into the model such as DP and federated learning. It is also possible to use TF Encrypted as the provider for TF-based encryption using PySyft, allowing for tighter integration between the 2 libraries. Similar to CrypTen and TF Encrypted, PySyft is a high-level library that attempts to make it easy for machine learning researchers to transition to build privacy-preserving models. However, PySyft should currently only be used as a research tool because many of its underlying protocols are not secure enough to be used with confidence.

CrypTFlow [281] differs from the aforementioned libraries in that it is a compiler for TF models rather than a programming interface. CrypTFlow takes a TF model as an input and outputs code that can run under an SMPC model. An advantage that CrypTFlow has over CrypTen, TF Encrypted, and PySyft is that, as part of its compilation process, CrypTFlow performs a number of optimization steps that in the other libraries would have to be done by hand or cannot be performed at all. For example, when converting floating-point numbers to a fixed-precision representation (which is necessary because

SMPC works inside a finite field), CrypTFlow chooses the smallest precision level that will match the classification accuracy of floating-point code. This, along with the other optimizations performed during the compilation process, means that it is possible to (efficiently) run much larger models in CrypTFlow than may be possible in other libraries. The possible real-world impact of CrypTFlow has been shown by running 2 networks designed for predicting lung disease from chest x-rays [282] and diabetic retinopathy [283] from retinal images. It is also possible to use CrypTFlow in conjunction with secure enclaves such as Software Guard Extension 41 (Intel Corporation) to work within the stricter malicious security assumptions; this is stricter than assuming semihonest parties because malicious parties may deviate from the defined protocol. The provision of malicious security means that CrypTFlow is more suitable for use in environments where extreme caution must be taken with the data set being used. Similar to CrypTen, the main issue with CrypTFlow is that it currently does not support the training of machine learning models because it is difficult to use the graphics processing unit in such a setting, meaning that there is still the need to be able to process plaintext data during the training process, which is not compatible with many of the scenarios where one may want to use privacy-preserving machine learning techniques.

An example of how SMPC protocols and SMPC-supporting machine learning libraries can be used is shown in the study by Hong et al [284], which used TF Encrypted to train a classifier on 2 genomic data sets, each containing a large number of features (12,634 and 17,814 features per sample), to detect tumors as part of the iDASH challenge. This task had an additional challenge because the 2 data sets were heavily imbalanced, but common countermeasures to this are difficult to implement in an SMPC framework. For example, resampling is commonly used to overcome this, but because the labels are private in SMPC, this is impossible. To overcome the imbalance, the weighting of samples in the loss function was adjusted to place a higher emphasis on those from the minority class. The study's best results had an accuracy of 69.48%, which is close to the classifier trained on the plaintext data, which showed an accuracy of 70%. This demonstrates that it is possible to train machine learning models on encrypted data; the study also noted that the TF Encrypted framework is easy to use for anyone familiar with TF, meaning that privacy-preserving machine learning is accessible to experts from both machine learning and cryptography fields.

CrypTen, TF Encrypted, and PySyft all have the advantage that they work closely with commonly used machine learning libraries (PyTorch, TF, and both PyTorch and TF, respectively), meaning that there is less of a learning curve required to make the existing models privacy preserving compared with tools such as CrypTFlow. This ease of use comes at the cost of efficiency, however, because more complex tools such as CrypTFlow are able to work at a lower level and perform more optimizations, allowing larger models to be encrypted.

Fully HE

HE is a type of encryption wherein the result of computations on the encrypted data, when decrypted, mirror the result of the same computations carried out on the plaintext data. Specifically, FHE is an encryption protocol that supports any computation on the ciphertext. Attempts have been made to apply FHE to machine learning [285,286]. Traditionally, because of the significant computational overhead required to run FHE computations, these models were trained in plaintext data; for example, it took 570 seconds to evaluate CryptoNet on the Modified National Institute of Standards and Technology data set [285]. It is only recently that we have been able to train a full classification model using FHE computations [36]. The main benefit of FHE over SMPC is that it does not require multiple and separate trusted parties; the models can be trained and run on encrypted data by a single entity. This makes FHE a more promising prospect than SMPC for problems involving data that are too sensitive to be entrusted to multiple parties (or in situations where multiple trusted parties may not be available).

Applying FHE to privacy-preserving machine learning is a relatively new area of research, and thus there are few tools that tie the 2 concepts together, with most research focusing on specific model implementations rather than on creating a general framework for FHE machine learning. One such tool, however, is CHET [287]. CHET is an optimizing compiler that takes a tensor circuit as an input and outputs an executable that can then be run with HE libraries such as Simple Encrypted Arithmetic Library (Microsoft Research) [288] or Homomorphic Encryption for Arithmetic of Approximate Numbers [289]. This automates many of the laborious processes (eg, encryption parameter setting) that are required when creating circuits that work with FHE libraries; these processes also require FHE domain knowledge, which we cannot expect many machine learning experts to possess. Hence, the use of CHET can result in more efficient FHE models. For example, the authors of CHET claim that it reduces the running time for analyzing a particular medical image model (provided by their industry partners) from 18 hours (the original, unoptimized FHE model) to just 5 minutes. However, despite CHET using numerous optimizing methods during its compilation phase, the resulting encrypted models are still restrictively slow (when compared with their nonencrypted counterparts). Not only does this mean it is only practical to use CHET with smaller models, but it also means that it is impractical to train a model using CHET. It is also important to consider whether FHE provides a level of security and privacy that is high enough for the task at hand; some current regulations argue that encryption is a form of pseudonymization rather than anonymization [290] because it is possible to decrypt encrypted data.

Legal Framework Support

Although general data protection laws such as GDPR and DPA and health care–specific guidelines have been proposed for a while, data anonymization practitioners still demand a combined and intuitive reference list to check. In this discussion, we tentatively construct a policy base by collecting and sorting the available guidance provided by 4 lawful aspects in an effort to benefit future intelligent data anonymization for health care.

The policy base was constructed by considering the documentation provided in accordance with legal frameworks

and guidelines proposed by government-accountable institutions, that is, the GDPR, particularly Article 5 [291]; the DPA [292]; the ICO (mainly based on the code of practice); and the NHS (with documents published in 2013 [293], 2015 [294], 2017 [75], 2019 [74,295], and 2021 [296,297]). Fundamentally, any organization (eg, the NHS or a UK company) that holds personal identifiable information is required to register with the ICO, and subsequently perform possible data anonymization followed by a reidentification risk assessment to evaluate the effectiveness of the anonymized data in line with the DPA (the UK implementation of the GDPR). In the case where the NHS or a UK company realizes that a data breach has occurred, it is required to report this to the ICO. In addition, the ICO provides guidance to help the NHS or UK companies to better understand the lawful basis for processing sensitive information. Recently, the ICO [298] and the European Data Protection Board [299] published their statements on the processing of personal identifiable data in coping with the COVID-19 outbreak.

From the NHS perspective, pseudonyms should be used on a one-off and consistent basis. In terms of the best practice recommendations, they recommend adopting cryptographic hash functions (eg, MD5, SHA-1, and SHA-2) to create a fixed-length hash. We argue that the encrypted data might be less useful for possible later data analysis and explainability research. We summarize the suggestions provided by the 4 aforementioned entities in Textbox 2.

Textbox 2. Guidance provided by 4 lawful aspects.

General Data Protection Regulation

- Accuracy
- Accountability
- Storage limitation
- Purpose limitation
- Data minimization
- Purpose limitation
- Lawfulness, fairness, and transparency

Data Protection Act

- Notify any personal data breach
- Settle system interruption or restoration
- Implement disclosure-risk measures
- Define legal basis for data processing
- Establish precise details of any processing
- Prevent unauthorized processing and inference
- Conduct data protection impact assessment
- Test anonymization effectiveness through reidentification
- No intent, threaten, or damage to cause in reidentification
- Ensure data integrity when malfunctions occur

Information Commissioner's Office

- Remove high-risk records
- Remove high-risk attribute
- Use average value of each group
- Use the week to replace the exact date
- Swap values of attributes with high risk
- Use partial postcode instead of full address
- Define a threshold and suppress the minority
- Probabilistically perturb categorical attributes
- Aggregate multiple variables into new classes
- Use city instead of postcode and house number, street
- Recode specific values into less-specific range
- Use secret key to link back (data owner only)
- Add noise to numerical data with low variations

National Health Service

- Round off the totals
- Swap data attributes
- Use identifier ranges
- Mask part of the data
- Use age rather than date of birth
- Change the sort sequence
- Use the first part of the postcode
- Remove direct identifiers (National Health Service number)

- Risk assessment of indirect identifiers
- Provide only a sample of the population
- Provide range banding rather than exact data
- If aggregate totals less than 5, use pseudonyms

Results Summary for RQs

Here we present the results of the 5 defined RQs (Textbox 3) and, in the next section, discuss 3 open questions in real-world EHR data anonymization. The overall results of this SLM study are summarized in Figure 7.

Textbox 3. Review questions.

- Review question (RQ) 1: Do best practices exist for the anonymization of realistic electronic health record (EHR) data?
  - As the leading question of this systematic literature mapping study, we answer this question by exploring the answers to the other 4 RQs. It is theoretically feasible but practically challenging. On the basis of the answers to the remaining 4 questions, theoretical operations, privacy models, reidentification risk, and usability measurements are sufficient. Despite this, anonymization is practically difficult mainly because of 2 reasons: (1) the knowledge gap between health care professionals and privacy law (usually requiring huge collaborative efforts by clinical science, law, and data science), although we have summarized all lawful bases in the following subsection; and (2) automatic anonymization of EHR data is nontrivial and very case dependent.
- RQ 2: What are the most frequently applied data anonymization operations, and how can these operations be applied?
  - We investigated 7 categories of basic data anonymization operations in 16 articles, most of which are summarized in Multimedia Appendix 3. Apart from their fundamental uses, they can also be incorporated into the data anonymization process in both conventional and machine learning–based privacy models.
- RQ 3: What are the existing conventional and machine learning–based privacy models for measuring the level of anonymity? Are they practically useful in handling real-world health care data? Are there any new trends?
  - We presented 40 conventional (a taxonomy for relational data is summarized as part of Figure 7) privacy models and 32 machine learning–based privacy models from a total of 104 articles (summarized as part of Table 1). From this, we have observed that combinations of a deep learning architecture and one or more data anonymization operations have become a trend, particularly techniques based on (conditional-) generative adversarial networks. We have also realized that despite the increasing number of publications from the computer vision community, they rarely use real-world sensitive medical data. For the applicability of existing privacy models, we present an ablation study (Multimedia Appendix 7 [181,300-303]) using publicly accessible EHRs in the next subsection as part of the discussion.
- RQ 4: What metrics could be adopted to measure the reidentification risk and usability of the anonymized data?
  - We investigated 7 (from 4 articles) and 15 (from 19 articles) metrics to quantify the risk of reidentification and degree of usability of the anonymized data. Measuring reidentification risk requires a pair of raw and anonymized data records in which the original data are treated as an object of reference and compared with the anonymized data in terms of statistical difference. Such a difference may not sufficiently reveal the true risk of reidentification. To further investigate this issue, we combined the privacy models for discussing the trade-offs between these 2 privacy aspects. In contrast, more usability metrics were proposed because of the wider availability of performance indicators.
- RQ 5: What are the off-the-shelf data anonymization tools based on conventional privacy models and machine learning?
  - We investigated and compared 19 data anonymization tools (reported in 36 articles), of which 15 are based on privacy models (compared in Tables 5 and 6), whereas the remaining 5 (compared in Table 7) rely on privacy-preserving machine learning (with issues summarized in the next subsection). However, there does not exist any off-the-shelf data anonymization tool that truly supports the current legal frameworks such as the General Data Protection Regulation and Data Protection Act to dispel the doubts and concerns of data owners (we filled this gap as well).

Discussion

Privacy-Usability Trade-offs and Practical Feasibility

The most important question to consider when data anonymization is required in the health care sector is the choice between the level of privacy and degree of usability. In Table 2, we listed parameter interval, which enables specific privacy model or models to be more practically configurable. The privacy level indicates the possible degree of privacy that can be achieved by each privacy model, where separate levels are provided for some variant models such as (α, k)-anonymity, stochastic t-closeness, and (ε, m)-anonymity. This problem can also be viewed as a trade-off between the risk of reidentification and data usability and can be quantified using specific methods [304-306].

It should be noted that the privacy models, reidentification risk measurements, and data usability metrics reviewed in this study are relatively easy to understand, with equations provided along with adequate descriptions. However, these concepts are difficult to deploy in real-world data anonymization tools. Even given the intensive investigations summarized above, the utility of

the anonymized data may not be easily measurable through a number of proposed metrics of reidentification risk and utility metrics.

Given this discrepancy observed from the ablation study we conducted (Multimedia Appendix 7), it is worth considering the problem domain when quantifying the reidentification risk as well as the utility of the anonymized data, although we summarized the existing measures in the previous section. Overall, the trade-offs between reidentification risk and usability are practically feasible yet problem dependent.

Issues of Privacy-Preserving Machine Learning

SMPC and FHE share some disadvantages. They both use encryption methods that work over finite fields, and thus they cannot natively work with floating-point numbers. All practical implementations instead use a fixed-precision representation, but this adds computational overhead, and the level of precision used can affect the accuracy of the results.

Another important issue is that of the trade-off between interpretability and privacy [307] which, where privacy-preserving machine learning is concerned, is highly skewed toward privacy; encrypted models are, because of their very nature, entirely black-box models. This is not only an issue in the health care field, where the explainability of machine learning models is an important issue [308], but also arguably in any machine learning application because of the GDPR's "right to an explanation" [309].

Encrypted, trained models are also still vulnerable to reverse-engineering attacks (regardless of the encryption method used) [278]; for example, a malicious user could use the outputs of a model to run a membership attack (ie, infer from the results of a model whether the input was from a member of the training set). Currently, the only known way to overcome this is to apply DP principles to the model, which adds yet another layer of complexity to the process. There are signs that existing libraries are starting to combat the possibility of such attacks by providing easy methods to apply DP to encrypted models; see, for example, the DP techniques available in PySyft in the *SMPC Frameworks* section above.

It is also important to remember that, as noted previously, any type of encryption is regarded as a form of pseudonymization rather than anonymization because the encrypted data can be decrypted by anyone with access to the encryption key. However, we note that much of the current guidance on viewing encryption techniques as anonymization or pseudonymization is ambiguous; for example, ICO guidance [290] suggests that encrypted data is classified as anonymized data so long as the party responsible for the encryption of the personal data is not also responsible for the processing of the encrypted data (because then the party processing the data would not be in possession of the encryption key and would therefore be unable to reverse the encryption). As such, it is important to carefully consider whether privacy-preserving machine learning techniques fully satisfy the requirements set out in law. For instance, tools that also include other privacy techniques, such as PySyft, may be more useful in situations where true anonymization is required.

Overall, privacy-preserving machine learning is a promising area of research, although more work needs to be undertaken to ensure that such methods are ready for use in industrial applications; many of the tools currently available are only suitable for research rather than practical application. There also needs to be some consideration over which privacy-preserving methods best suit the needs of the application. SMPC currently offers a more viable approach than FHE because of its ability to run (and, more importantly, train) larger models, although the need to have multiple trusted parties may mean that it is seen as less secure than FHE. Meanwhile, FHE for privacy-preserving machine learning is still an emerging field, and it is encouraging to see research being undertaken by both the machine learning and cryptographic communities to improve the practicality of FHE methods by improving the running time of encrypted models and reducing the level of cryptographic knowledge needed to create efficient, encrypted models using FHE.

Conclusions

In this SLM study, we presented a comprehensive overview of data anonymization research for health care by investigating both conventional and emerging privacy-preserving techniques. Given the results and the discussions regarding the 5 proposed RQs, privacy-preserving data anonymization for health care is a promising domain, although more studies are required to be conducted to ensure more reliable industrial applications.

Acknowledgments

This study was sponsored by the UK Research and Innovation fund (project 312409) and Cievert Ltd.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Math notation system.
[PDF File (Adobe PDF File), 96 KB - medinform_v9i10e29871_app1.pdf]

Multimedia Appendix 2
Typical direct identifiers and quasi-identifiers in UK electronic health record data.

[PDF File (Adobe PDF File), 84 KB - medinform_v9i10e29871_app2.pdf]

Multimedia Appendix 3
Examples of fundamental data anonymization operations.
[PDF File (Adobe PDF File), 68 KB - medinform_v9i10e29871_app3.pdf]

Multimedia Appendix 4
Conventional privacy models.
[PDF File (Adobe PDF File), 321 KB - medinform_v9i10e29871_app4.pdf]

Multimedia Appendix 5
Usability metrics for privacy models.
[PDF File (Adobe PDF File), 190 KB - medinform_v9i10e29871_app5.pdf]

Multimedia Appendix 6
Reidentification risk metrics for privacy models.
[PDF File (Adobe PDF File), 139 KB - medinform_v9i10e29871_app6.pdf]

Multimedia Appendix 7
Ablation study for privacy-usability trade-offs and practical feasibility.
[PDF File (Adobe PDF File), 195 KB - medinform_v9i10e29871_app7.pdf]

References

1. Duggal R, Brindle I, Bagenal J. Digital healthcare: regulating the revolution. Br Med J 2018 Jan 15;360:k6. [doi: 10.1136/bmj.k6] [Medline: 29335296]
2. Li Z, Wang C, Han M, Xue Y, Wei W, LI LJ, et al. Thoracic disease identification and localization with limited supervision. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; Jun 18-23, 2018; Salt Lake City, UT, USA. [doi: 10.1109/cvpr.2018.00865]
3. Wang G, Zuluaga MA, Li W, Pratt R, Patel PA, Aertsen M, et al. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE Trans Pattern Anal Mach Intell 2019 Jul 1;41(7):1559-1572. [doi: 10.1109/tpami.2018.2840695]
4. Zuo Z, Yang L, Peng Y, Chao F, Qu Y. Gaze-informed egocentric action recognition for memory aid systems. IEEE Access 2018 Mar 1;6:12894-12904. [doi: 10.1109/access.2018.2808486]
5. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020 Jan;577(7792):706-710. [doi: 10.1038/s41586-019-1923-7] [Medline: 31942072]
6. Veličković P, Karazija L, Lane N, Bhattacharya S, Liberis E, Liò P, et al. Cross-modal recurrent models for weight objective prediction from multimodal time-series data. In: Proceedings of the PervasiveHealth '18: 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: PervasiveHealth '18:12th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 21-24, 2018; New York. [doi: 10.1145/3240925.3240937]
7. Zhelezniak V, Savkov A, Shen A, Moramarco F, Flann J, Hammerla N. Don't settle for average, go for the max: fuzzy sets and max-pooled word vectors. arXiv. 2019. URL: https://arxiv.org/abs/1904.13264 [accessed 2021-08-30]
8. Huang SY, Omkar, Yoshida Y, Inda AJ, Xavier CX, Mu WC, et al. Microstrip line-based glucose sensor for noninvasive continuous monitoring using the main field for sensing and multivariable crosschecking. IEEE Sensors J 2019 Jan 15;19(2):535-547. [doi: 10.1109/jsen.2018.2877691]
9. Kaisti M, Tadi MJ, Lahdenoja O, Hurnanen T, Saraste A, Pankaala M, et al. Stand-alone heartbeat detection in multidimensional mechanocardiograms. IEEE Sensors J 2019 Jan 1;19(1):234-242. [doi: 10.1109/jsen.2018.2874706]
10. Iacobucci G. Row over Babylon's chatbot shows lack of regulation. Br Med J 2020 Feb 28;368:m815. [doi: 10.1136/bmj.m815] [Medline: 32111647]
11. Spencer T, Noyes E, Biederman J. Telemedicine in the management of ADHD: literature review of telemedicine in ADHD. J Atten Disord 2020 Jan;24(1):3-9. [doi: 10.1177/1087054719859081] [Medline: 31257978]
12. Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank: current status and what it means for epidemiology. Health Policy Technol 2012 Sep;1(3):123-126. [doi: 10.1016/j.hlpt.2012.07.003]
13. Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 2018 Oct;562(7726):210-216 [FREE Full text] [doi: 10.1038/s41586-018-0571-7] [Medline: 30305740]

14. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, 100 000 Genomes Project. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. Br Med J 2018 Apr 24;361:k1687. [doi: 10.1136/bmj.k1687] [Medline: 29691228]
15. Heath I. Boost for sustainable healthcare. Br Med J 2020 Jan 28;368:m284. [doi: 10.1136/bmj.m284] [Medline: 31992564]
16. Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity (Edinb) 2020 Apr;124(4):525-534 [FREE Full text] [doi: 10.1038/s41437-020-0303-2] [Medline: 32139886]
17. Moberly T. Should we be worried about the NHS selling patient data? Br Med J 2020 Jan 15;368:m113. [doi: 10.1136/bmj.m113] [Medline: 31941645]
18. Human Rights Act 1998. Legislation - UK Public General Acts. 1998. URL: https://www.legislation.gov.uk/ukpga/1998/42/schedule/1 [accessed 2021-08-30]
19. Wang Z, Vineet V, Pittaluga F, Sinha S, Cossairt O, Bing KS. Privacy-preserving action recognition using coded aperture videos. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019 Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 16-17, 2019; Long Beach, CA, USA. [doi: 10.1109/cvprw.2019.00007]
20. Speciale P, Schonberger J, Sinha S, Pollefeys M. Privacy preserving image queries for camera localization. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019 Presented at: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 27- Nov 2, 2019; Seoul, Korea (South). [doi: 10.1109/iccv.2019.00157]
21. Li J, Khodak M, Caldas S, Talwalkar A. Differentially private meta-learning. arXiv. 2019. URL: https://arxiv.org/abs/1909.05830 [accessed 2021-08-30]
22. Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data. European Union Agency for Cybersecurity. 1980. URL: https://www.enisa.europa.eu/topics/threat-risk-management/risk-management/current-risk/laws-regulation/data-protection-privacy/oecd-recommendation-of-the-council [accessed 2021-08-30]
23. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Legislation - Regulations Originating from the EU. 2016. URL: https://www.legislation.gov.uk/eur/2016/679/contents# [accessed 2021-08-30]
24. The Lancet. Artificial intelligence in health care: within touching distance. Lancet 2017 Dec 23;390(10114):2739. [doi: 10.1016/S0140-6736(17)31540-4] [Medline: 29303711]
25. Iacobucci G. Patient data were shared with Google on an "inappropriate legal basis," says NHS data guardian. Br Med J 2017 May 18;357:j2439. [doi: 10.1136/bmj.j2439] [Medline: 28522583]
26. Lee N. The Lancet Technology: November, 2014. Trouble on the radar. Lancet 2014 Nov 29;384(9958):1917. [doi: 10.1016/s0140-6736(14)62267-4] [Medline: 25478615]
27. Shah H. The DeepMind debacle demands dialogue on data. Nature 2017 Jul 19;547(7663):259. [doi: 10.1038/547259a] [Medline: 28726841]
28. Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, Davidson KW, et al. Using EHRs to integrate research with patient care: promises and challenges. J Am Med Inform Assoc 2012;19(5):684-687 [FREE Full text] [doi: 10.1136/amiajnl-2012-000878] [Medline: 22542813]
29. Evans R. Samaritans radar app. Nurs Stand 2014 Dec 15;29(15):33. [doi: 10.7748/ns.29.15.33.s40] [Medline: 25492781]
30. Thompson CL, Morgan HM. Ethical barriers to artificial intelligence in the national health service, United Kingdom of Great Britain and Northern Ireland. Bull World Health Organ 2020 Apr 01;98(4):293-295 [FREE Full text] [doi: 10.2471/BLT.19.237230] [Medline: 32284657]
31. Hawkes N. NHS data sharing deal with Google prompts concern. Br Med J 2016 May 05;353:i2573. [doi: 10.1136/bmj.i2573] [Medline: 27150956]
32. Royal Free - Google DeepMind trial failed to comply with data protection law. Information Commissioner's Office. 2017. URL: https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/ [accessed 2021-08-30]
33. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. J Med Internet Res 2019 May 31;21(5):e13484 [FREE Full text] [doi: 10.2196/13484] [Medline: 31152528]
34. Elger BS, Iavindrasana J, Lo Iacono L, Müller H, Roduit N, Summers P, et al. Strategies for health data exchange for secondary, cross-institutional clinical research. Comput Methods Programs Biomed 2010 Sep;99(3):230-251. [doi: 10.1016/j.cmpb.2009.12.001] [Medline: 20089327]
35. Annas GJ. HIPAA regulations - a new era of medical-record privacy? N Engl J Med 2003 Apr 10;348(15):1486-1490. [doi: 10.1056/NEJMlim035027] [Medline: 12686707]
36. Badawi A, Chao J, Lin J, Mun CF, Jie SJ, Tan BH, et al. Towards the AlexNet moment for homomorphic encryption: HCNN, the first homomorphic CNN on encrypted data with GPUs. arXiv. 2018. URL: https://arxiv.org/abs/1811.00778 [accessed 2021-08-30]

37. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). World Health Organization. 2020. URL: https://tinyurl.com/5bcfwe8a [accessed 2021-08-30]
38. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020 May;20(5):533-534 [FREE Full text] [doi: 10.1016/S1473-3099(20)30120-1] [Medline: 32087114]
39. COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE). Arcgis. URL: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 [accessed 2021-08-30]
40. UK summary : The official UK government website for data and insights on coronavirus (COVID-19). GOV.UK. 2021. URL: https://coronavirus.data.gov.uk/ [accessed 2021-08-30]
41. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
42. Flaxman S, Mishra S, Gandy A, Unwin H, Mellan TA, Coupland H, Imperial College COVID-19 Response Team, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature 2020 Aug;584(7820):257-261. [doi: 10.1038/s41586-020-2405-7] [Medline: 32512579]
43. Wu J, Leung K, Leung G. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 2020 Feb 29;395(10225):689-697 [FREE Full text] [doi: 10.1016/S0140-6736(20)30260-9] [Medline: 32014114]
44. Leung GM, Leung K. Crowdsourcing data to mitigate epidemics. Lancet Digit Health 2020 Apr;2(4):156-157 [FREE Full text] [doi: 10.1016/S2589-7500(20)30055-8] [Medline: 32296776]
45. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. Lancet Digit Health 2020 Apr;2(4):201-208 [FREE Full text] [doi: 10.1016/S2589-7500(20)30026-1] [Medline: 32309796]
46. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science 2020 May 08;368(6491):eabb6936 [FREE Full text] [doi: 10.1126/science.abb6936] [Medline: 32234805]
47. Coronavirus: UK considers virus-tracing app to ease lockdown. BBC News. 2020. URL: https://www.bbc.co.uk/news/technology-52095331 [accessed 2021-08-30]
48. Kuhn C, Beck M, Strufe T. Covid notions: towards formal definitions - and documented understanding - of privacy goals and claimed protection in proximity-tracing services. Online Soc Netw Media 2021 Mar;22:100125 [FREE Full text] [doi: 10.1016/j.osnem.2021.100125] [Medline: 33681543]
49. Coronavirus: Moscow rolls out patient-tracking app. BBC News. 2020. URL: https://www.bbc.co.uk/news/technology-52121264 [accessed 2021-08-30]
50. The Lancet Digital Health. Reflecting on a future ready for digital health. Lancet Digit Health 2020 May;2(5):e209 [FREE Full text] [doi: 10.1016/S2589-7500(20)30087-X] [Medline: 32373784]
51. Nay O. Can a virus undermine human rights? Lancet Public Health 2020 May;5(5):238-239 [FREE Full text] [doi: 10.1016/S2468-2667(20)30092-X] [Medline: 32325013]
52. Chen S, Yang J, Yang W, Wang C, Bärnighausen T. COVID-19 control in China during mass population movements at New Year. Lancet 2020 Mar 07;395(10226):764-766 [FREE Full text] [doi: 10.1016/S0140-6736(20)30421-9] [Medline: 32105609]
53. Zhao S, Lin Q, Ran J, Musa S, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: a data-driven analysis in the early phase of the outbreak. Int J Infect Dis 2020 Mar;92:214-217 [FREE Full text] [doi: 10.1016/j.ijid.2020.01.050] [Medline: 32007643]
54. NHS COVID-19 app support. NHS. 2020. URL: https://www.covid19.nhs.uk/ [accessed 2021-08-30]
55. Beigi G, Liu H. A survey on privacy in social media: identification, mitigation, and applications. ACM/IMS Trans Data Sci 2020 Feb;1(1):1-38. [doi: 10.1145/3343038]
56. Fletcher S, Islam MZ. Decision tree classification with differential privacy. ACM Comput Surv 2019 Sep;52(4):1-33. [doi: 10.1145/3337064]
57. Fung BC, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. ACM Comput Surv 2010 Jun;42(4):1-53. [doi: 10.1145/1749603.1749605]
58. Wagner I, Eckhoff D. Technical privacy metrics. ACM Comput Surv 2018 Jul;51(3):1-38. [doi: 10.1145/3168389]
59. GDPR personal data. Intersoft Consulting. URL: https://gdpr-info.eu/issues/personal-data/ [accessed 2021-08-30]
60. Art. 9 GDPR Processing of special categories of personal data. Intersoft Consulting. URL: https://gdpr-info.eu/art-9-gdpr/ [accessed 2021-08-30]
61. Emam K. Guide to the De-Identification of Personal Health Information. Boca Raton, Fl: Auerbach Publications; 2013.
62. Zigomitros A, Casino F, Solanas A, Patsakis C. A survey on privacy properties for data publishing of relational data. IEEE Access 2020 Mar 11;8:51071-51099. [doi: 10.1109/access.2020.2980235]
63. Wang J, Zhou S, Wu J, Liu C. A new approach for anonymizing relational and transaction data. In: Proceedings of the 2nd International Conference on Healthcare Science and Engineering. ICHSE 2018. Singapore: Springer; 2019.

64. Amiri F, Yazdani N, Shakery A. Bottom-up sequential anonymization in the presence of adversary knowledge. Inf Sci: Int J 2018 Jun;450:316-335 [FREE Full text] [doi: 10.1016/j.ins.2018.03.027]
65. Gao S, Ma J, Sun C, Li X. Balancing trajectory privacy and data utility using a personalized anonymization model. J Netw Comput Appl 2014 Feb;38:125-134 [FREE Full text] [doi: 10.1016/j.jnca.2013.03.010]
66. Mortazavi R, Erfani S. GRAM: an efficient (k, l) graph anonymization method. Expert Syst Appl 2020 Sep 1;153:113454 [FREE Full text] [doi: 10.1016/j.eswa.2020.113454]
67. Ninggal M, Abawajy J. Utility-aware social network graph anonymization. J Netw Comput Appl 2015 Oct;56:137-148 [FREE Full text] [doi: 10.1016/j.jnca.2015.05.013]
68. Stallings W. Information Privacy Engineering and Privacy by Design: Understanding Privacy Threats, Technology, and Regulations Based on Standards and Best Practices. Boston, MA: Addison-Wesley Professional; 2019.
69. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Br Med J 2010 Jan 28;340:c181 [FREE Full text] [doi: 10.1136/bmj.c181] [Medline: 20110312]
70. Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, et al. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med Res Methodol 2016 Jul 08;16 Suppl 1(Suppl 1):77 [FREE Full text] [doi: 10.1186/s12874-016-0169-4] [Medline: 27410040]
71. Recommendation CM/Rec(2019)2 of the Committee of Ministers to member States on the protection of health-related data. Council of Europe. 2019. URL: https://www.apda.ad/sites/default/files/2019-03/CM_Rec%282019%292E_EN.pdf [accessed 2021-08-30]
72. HIPAA PHI: definition of PHI and list of 18 identifiers. UC Berkeley Human Research Protection Program. URL: https://cphs.berkeley.edu/hipaa/hipaa18.html [accessed 2021-08-30]
73. Information sharing policy. NHS. 2019. URL: https://www.england.nhs.uk/wp-content/uploads/2019/10/information-sharing-policy-v4.1.pdf [accessed 2021-08-30]
74. Pseudonymisation policy. Kernow NHS Foundation Trust. 2019. URL: http://policies.kernowccg.nhs.uk/DocumentsLibrary/KernowCCG/ManagingInformation/Policies/PseudonymisationPolicy.pdf [accessed 2021-08-30]
75. Anonymisation of data (Pseudonymisation) policy and procedure. Solent NHS Foundation Trust. 2020. URL: https://www.solent.nhs.uk/media/1262/pseudonymisation-policy.pdf [accessed 2021-08-30]
76. Recital 26 not applicable to anonymous data. Intersoft Consulting. URL: https://gdpr-info.eu/recitals/no-26/ [accessed 2021-08-30]
77. What is personal data? Information Commissioner's Office. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/ [accessed 2021-08-30]
78. Templ M. Anonymization and re-identification risk of personal data. In: Proceedings of the Gästekolloquium Psychologisches Institut Universität Zürich. 2020 Presented at: Gästekolloquium Psychologisches Institut Universität Zürich; Nov 9, 2020; Zurich. [doi: 10.4414/saez.2019.17441]
79. Bandara P, Bandara H, Fernando S. Evaluation of re-identification risks in data anonymization techniques based on population uniqueness. In: Proceedings of the 2020 5th International Conference on Information Technology Research (ICITR). 2020 Presented at: 2020 5th International Conference on Information Technology Research (ICITR); Dec 2-4, 2020; Moratuwa, Sri Lanka. [doi: 10.1109/icitr51448.2020.9310884]
80. Garousi V, Bauer S, Felderer M. NLP-assisted software testing: a systematic mapping of the literature. Inf Softw Technol 2020 Oct;126:106321 [FREE Full text] [doi: 10.1016/j.infsof.2020.106321]
81. Petersen K, Feldt R, Mujtaba S, Mattsson M. Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE). 2008 Presented at: 12th International Conference on Evaluation and Assessment in Software Engineering (EASE); Jun 26-17, 2008; Italy p. 68-77. [doi: 10.5555/2227115.2227123]
82. Witti M, Konstantas D. IOT and security-privacy concerns: a systematic mapping study. Int J Netw Secur Appl 2018 Nov 21;10(6):3319816. [doi: 10.2139/ssrn.3319816]
83. Budgen D, Brereton P, Williams N, Drummond S. What support do systematic reviews provide for evidence-informed teaching about software engineering practice? e-Infor Softw Eng J 2020;14(1):7-60. [doi: 10.37190/e-inf200101]
84. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 2015 Aug;64:1-18. [doi: 10.1016/j.infsof.2015.03.007]
85. Cochrane database of systematic reviews. Cochrane Library. URL: https://www.cochranelibrary.com/cdsr/about-cdsr [accessed 2021-08-30]
86. Centre for reviews and dissemination. University of York. URL: https://www.york.ac.uk/crd/ [accessed 2021-08-30]
87. Health technology assessment. National Institute for Health Research. URL: https://www.nihr.ac.uk/explore-nihr/funding-programmes/health-technology-assessment.htm [accessed 2021-08-30]
88. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. 2014 Presented at: EASE '14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering; May 13-14, 2014; London England United Kingdom. [doi: 10.1145/2601248.2601268]

89. Paez A. Gray literature: an important resource in systematic reviews. J Evid Based Med 2017 Aug;10(3):233-240. [doi: 10.1111/jebm.12266] [Medline: 28857505]
90. Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev 2007 Apr 18(2):MR000010. [doi: 10.1002/14651858.MR000010.pub3] [Medline: 17443631]
91. Saleh AA, Ratajeski MA, Bertolet M. Grey literature searching for health sciences systematic reviews: a prospective study of time spent and resources utilized. Evid Based Libr Inf Pract 2014;9(3):28-50 [FREE Full text] [doi: 10.18438/b8dw3k] [Medline: 25914722]
92. Nimmer RT, Krauthaus PA. Information as a commodity: new imperatives of commercial law. Law Contemp Probs 1992;55(3):103-130. [doi: 10.2307/1191865]
93. Adams RJ, Smart P, Huff AS. Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies. Int J Manag Rev 2016 Apr 19;19(4):432-454. [doi: 10.1111/ijmr.12102]
94. Chen C, Li L, Wu B, Hong C, Wang L, Zhou J. Secure social recommendation based on secret sharing. In: Proceedings of the 24th European Conference on Artificial Intelligence - ECAI 2020. 2020 Presented at: 24th European Conference on Artificial Intelligence - ECAI 2020; Aug 31- Sep 4, 2020; Spain URL: https://ecai2020.eu/papers/609_paper.pdf
95. Choi E, Biswal S, Malin B, Duke J, Stewart W, Sun J. Generating multi-label discrete patient records using generative adversarial networks. In: Proceedings of the 2nd Machine Learning for Healthcare Conference. 2017 Aug Presented at: Proceedings of the 2nd Machine Learning for Healthcare Conference; Aug 18-19, 2017; Boston, Massachusetts, USA URL: http://proceedings.mlr.press/v68/choi17a.html
96. Hjelm R, Jacob A, Trischler A, Che G, Cho K, Bengio Y. Boundary-seeking generative adversarial networks. arXiv. 2017. URL: https://arxiv.org/abs/1702.08431 [accessed 2021-08-30]
97. Chen C, Zhou J, Wu B, Fang W, Wang L, Qi Y, et al. Practical privacy preserving POI recommendation. ACM Trans Intell Syst Technol 2020 Sep 05;11(5):1-20. [doi: 10.1145/3394138]
98. Where the world builds software. GitHub. URL: https://github.com/ [accessed 2021-08-30]
99. Anonymisation: managing data protection risk code of practice. Information Commissioner's Office. URL: https://ico.org.uk/media/1061/anonymisation-code.pdf [accessed 2021-08-30]
100. Defays D, Anwar M. Masking microdata using micro-aggregation. J Off Stat 1998;14(4):449-461 [FREE Full text]
101. Fienberg S, McIntyre J. Data swapping: variations on a theme by Dalenius and Reiss. J Off Stat 2005;21(2):309 [FREE Full text] [doi: 10.1007/978-3-540-25955-8_2]
102. Li D, He X, Cao L, Chen H. Permutation anonymization. J Intell Inf Syst 2015 Aug 4;47(3):427-445. [doi: 10.1007/s10844-015-0373-4]
103. Nin J, Herranz J, Torra V. Rethinking rank swapping to decrease disclosure risk. Data Knowl Eng 2008 Jan;64(1):346-364. [doi: 10.1016/j.datak.2007.07.006]
104. Gouweleeuw J, Kooiman P, Willenborg L, Wolf P. Post randomisation for statistical disclosure control: theory and implementation. J Off Stat 1998;14(4):463-478. [doi: 10.1002/9781118348239]
105. Brand R. Microdata protection through noise addition. In: Inference Control in Statistical Databases. Berlin, Heidelberg: Springer; 2002.
106. Domingo-Ferrer J, Mateo-Sanz J. Resampling for statistical confidentiality in contingency tables. Comput Math Appl 1999 Dec;38(11-12):13-32. [doi: 10.1016/s0898-1221(99)00281-3]
107. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. Br Med J 2015 Mar 20;350:h1139 [FREE Full text] [doi: 10.1136/bmj.h1139] [Medline: 25794882]
108. Vijayarani S, Tamilarasi A. An efficient masking technique for sensitive data protection. In: Proceedings of the 2011 International Conference on Recent Trends in Information Technology (ICRTIT).: Jun 3-5, 2011; 2011 Presented at: 2011 International Conference on Recent Trends in Information Technology (ICRTIT); 2011; Chennai, India. [doi: 10.1109/icrtit.2011.5972275]
109. Dwork C. Differential privacy. In: Proceedings of International Colloquium on Automata, Languages, and Programming. 2011 Presented at: International Colloquium on Automata, Languages, and Programming; July 4-8, 2011; Zurich, Switzerland. [doi: 10.1007/978-1-4419-5906-5_752]
110. Kelly JP, Golden BL, Assad AA. Cell suppression: disclosure protection for sensitive tabular data. Netw Int J 1992 Jul;22(4):397-417. [doi: 10.1002/net.3230220407]
111. Samarati P, Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. CiteSeer - Technical Report, SRI International. 1998. URL: http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=6175EF3A3D18C80FE50ADB6B3E03B675?doi=10.1.1.37.5829 [accessed 2021-08-30]
112. Sweeney L. k-Anonymity: a model for protecting privacy. Int J Unc Fuzz Knowl Based Syst 2002;10(05):557-570. [doi: 10.1142/s0218488502001648]
113. Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-Anonymity through microaggregation. Data Min Knowl Disc 2005 Aug 23;11(2):195-212. [doi: 10.1007/s10618-005-0007-5]
114. Meyerson A, Williams R. On the complexity of optimal K-anonymity. In: Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2004 Presented at: PODS '04: Proceedings

of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems; Jun 14-16, 2004; Paris France. [doi: 10.1145/1055558.1055591]
115. Sweeney L. Achieving k-Anonymity privacy protection using generalization and suppression. Int J Unc Fuzz Knowl Based Syst 2012 May 02;10(05):571-588. [doi: 10.1142/s021848850200165x]
116. LeFevre K, DeWitt D, Ramakrishnan R. Incognito: efficient full-domain K-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. 2021 Aug 30 Presented at: SIGMOD '05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data; Jun 14-16, 2005; Baltimore Maryland. [doi: 10.1145/1066157.1066164]
117. LeFevre K, DeWitt D, Ramakrishnan R. Mondrian multidimensional K-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06). 2006 Presented at: 22nd International Conference on Data Engineering (ICDE'06); Apr 3-7, 2006; Atlanta, GA, USA. [doi: 10.1109/icde.2006.101]
118. Li J, Tao Y, Xiao X. Preservation of proximity privacy in publishing numerical sensitive data. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008 Presented at: SIGMOD '08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data; Jun 9-12, 2008; Vancouver Canada. [doi: 10.1145/1376616.1376666]
119. Chaurasia SK, Mishra N, Sharma S. Comparison of K-automorphism and K2-degree anonymization for privacy preserving in social network. Int J Comput Appl 2013 Oct;79(14):30-36. [doi: 10.5120/13811-1871]
120. Liu J, Wang K. On optimal anonymization for l+-diversity. In: Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010). 2010 Presented at: 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010); Mar 1-6, 2010; Long Beach, CA, USA. [doi: 10.1109/ICDE.2010.5447898]
121. Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering. 2007 Presented at: 2007 IEEE 23rd International Conference on Data Engineering; Apr 15-20, 2007; Istanbul, Turkey. [doi: 10.1109/ICDE.2007.367856]
122. Soria-Comas J, Domingo-Ferrer J, Sanchez D, Martinez S. t-closeness through microaggregation: strict privacy with enhanced utility preservation. In: Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering (ICDE). 2016 Presented at: 2016 IEEE 32nd International Conference on Data Engineering (ICDE); May 16-20, 2016; Helsinki, Finland. [doi: 10.1109/ICDE.2016.7498376]
123. Domingo-Ferrer J, Soria-Comas J. From t-closeness to differential privacy and vice versa in data anonymization. Knowl Based Syst 2015 Jan;74:151-158. [doi: 10.1016/j.knosys.2014.11.011]
124. Chawla S, Dwork C, McSherry F, Smith A, Wee H. Toward privacy in public databases. In: Theory of Cryptography. Berlin, Heidelberg: Springer; 2005.
125. Cao J, Karras P. Publishing microdata with a robust privacy guarantee. Proc VLDB Endowment 2012 Jul;5(11):1388-1399. [doi: 10.14778/2350229.2350255]
126. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography. Berlin, Heidelberg: Springer; 2006.
127. Hsu J, Gaboardi M, Haeberlen A, Khanna S, Narayan A, Pierce B. Differential privacy: an economic method for choosing epsilon. In: Proceedings of the 2014 IEEE 27th Computer Security Foundations Symposium. 2014 Presented at: 2014 IEEE 27th Computer Security Foundations Symposium; Jul 19-22, 2014; Vienna, Austria. [doi: 10.1109/csf.2014.35]
128. Krehbiel S. Choosing epsilon for privacy as a service. Proc Priv Enhanc Technol 2018;2019(1):192-205. [doi: 10.2478/popets-2019-0011]
129. Katewa V, Pasqualetti F, Gupta V. On the role of cooperation in private multi-agent systems. In: Privacy in Dynamical Systems. Singapore: Springer; 2020.
130. Holohan N, Antonatos S, Braghin S, Aonghusa P. ($ k $, $\epsilon $)-anonymity: $ k $-anonymity with $\epsilon $-differential privacy. Data Privacy @ IBMRisk and Privacy. 2017. URL: https://www.researchgate.net/publication/320223744_kepsilon-Anonymity_k-Anonymity_with_epsilon-Differential_Privacy [accessed 2021-08-30]
131. Holohan N, Braghin S, Mac AP, Levacher K. Diffprivlib: the IBM differential privacy library. arXiv. 2019. URL: https://arxiv.org/abs/1907.02444 [accessed 2021-08-30]
132. Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. Our data, ourselves: privacy via distributed noise generation. In: Advances in Cryptology - EUROCRYPT 2006. Berlin, Heidelberg: Springer; 2006.
133. Bild R, Kuhn K, Prasser F. SafePub: a truthful data anonymization algorithm with strong privacy guarantees. Proc Priv Enhanc Technol 2018;2018(1):67-87. [doi: 10.1515/popets-2018-0004]
134. Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy. J Asso Comput Machin 2013 Apr;60(2):1-25. [doi: 10.1145/2450142.2450148]
135. Beimel A, Nissim K, Stemmer U. Private learning and sanitization: pure vs. Approximate differential privacy. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Berlin, Heidelberg: Springer; 2013.
136. Dwork C, Roth A. The algorithmic foundations of differential privacy. In: Foundations and Trends in Theoretical Computer. Boston: Now Publishers Inc; 2014:211-407.

137. Abadi M, Chu A, Goodfellow I, McMahan H, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016 Presented at: CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; Oct 24-28, 2016; Vienna Austria. [doi: 10.1145/2976749.2978318]
138. Shi E, Chan TH, Rieffel E, Chow R, Song D. Privacy-preserving aggregation of time-series data. In: Proceedings of the Network and Distributed System Security Symposium. 2011 Presented at: Network and Distributed System Security Symposium; Feb 6-9, 2011; San Diego, California, USA URL: https://www.researchgate.net/publication/221655415_Privacy-Preserving_Aggregation_of_Time-Series_Data
139. Jelasity M, Birman K. Distributional differential privacy for large-scale smart metering. In: Proceedings of the 2nd ACM workshop on Information Hiding and Multimedia Security. 2014 Presented at: IH&MMSec '14: 2nd ACM workshop on Information Hiding and Multimedia Security; Jun 11-13, 2014; Salzburg Austria. [doi: 10.1145/2600918.2600919]
140. Chatzikokolakis K, Andrés M, Bordenabe N, Palamidessi C. Broadening the scope of differential privacy using metrics. In: Proceedings of the International Symposium on Privacy Enhancing Technologies Symposium. 2013 Presented at: International Symposium on Privacy Enhancing Technologies Symposium; Jul 10-12, 2013; Bloomington, IN, USA. [doi: 10.1007/978-3-642-39077-7_5]
141. Wang K, Fung B. Anonymizing sequential releases. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006 Presented at: KDD '06: 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 20-23, 2006; Philadelphia PA USA. [doi: 10.1145/1150402.1150449]
142. Oliveira SR, Zaiane OR. Privacy preserving clustering by data transformation. J Inf Data Manag 2010;1(1):37 [FREE Full text]
143. Brickell J, Shmatikov V. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008 Presented at: KDD '08: 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 24-27, 2008; Las Vegas Nevada USA. [doi: 10.1145/1401890.1401904]
144. Evfimievski A, Srikant R, Agrawal R, Gehrke J. Privacy preserving mining of association rules. Inf Syst 2004 Jun;29(4):343-364. [doi: 10.1016/j.is.2003.09.001]
145. Rastogi V, Suciu D, Hong S. The boundary between privacy and utility in data publishing. In: Proceedings of the 33rd International Conference on Very Large Data Bases. 2007 Presented at: VLDB '07: 33rd International Conference on Very Large Data Bases; Sep 23-27, 2007; Vienna, Austria p. 531-542. [doi: 10.5555/1325851.1325913]
146. Nergiz M, Atzori M, Clifton C. Hiding the presence of individuals from shared databases. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. 2007 Presented at: SIGMOD '07: 2007 ACM SIGMOD International Conference on Management of Data; Jun 11-14, 2007; Beijing China. [doi: 10.1145/1247480.1247554]
147. Rocher L, Hendrickx JM, de Montjoye Y. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 2019 Jul 23;10(1):3069 [FREE Full text] [doi: 10.1038/s41467-019-10933-3] [Medline: 31337762]
148. Hoshino N. Applying Pitman's sampling formula to microdata disclosure risk assessment. J Off Stat 2001;17(4):499-520. [doi: 10.1007/978-3-319-50272-4_3]
149. Zayatz L. Estimation of the percent of unique population elements on a microdata file using the sample. Bureau of the Census Statistical Research Division Report Series. 1991. URL: https://www.census.gov/srd/papers/pdf/rr91-08.pdf [accessed 2021-08-30]
150. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata. J Off Stat 1998;14(1):79-95.
151. Genz A, Bretz F. Computation of Multivariate Normal and t Probabilities. Heidelberg: Springer; 2009.
152. Wan Z, Vorobeychik Y, Xia W, Clayton EW, Kantarcioglu M, Ganta R, et al. A game theoretic framework for analyzing re-identification risk. PLoS One 2015 Mar 25;10(3):e0120592 [FREE Full text] [doi: 10.1371/journal.pone.0120592] [Medline: 25807380]
153. Gkountouna O, Angeli S, Zigomitros A, Terrovitis M, Vassiliou Y. km-Anonymity for continuous data using dynamic hierarchies. In: Privacy in Statistical Databases. Cham: Springer; 2014.
154. Poulis G, Loukides G, Gkoulalas-Divanis A, Skiadopoulos S. Anonymizing data with relational and transaction attributes. In: Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer; 2013.
155. Liu K, Terzi E. Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. 2008 Presented at: SIGMOD '08: 2008 ACM SIGMOD International Conference on Management of Data; Jun 9-12, 2008; Vancouver Canada. [doi: 10.1145/1376616.1376629]
156. Tai C, Yu P, Yang DN, Chen MS. Privacy-preserving social network publication against friendship attacks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011 Presented at: KDD '11: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 21-24, 2011; San Diego California. [doi: 10.1145/2020408.2020599]
157. Zou L, Chen L, Özsu MT. k-automorphism: a general framework for privacy preserving network publication. Proc VLDB Endow 2009 Aug;2(1):946-957. [doi: 10.14778/1687627.1687734]

158. Korolova A, Motwani R, Nabar S, Xu Y. Link privacy in social networks. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008 Presented at: CIKM '08: 17th ACM Conference on Information and Knowledge Management; Oct 26-30, 2008; Napa Valley California USA. [doi: 10.1145/1458082.1458123]
159. Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. 2008 Presented at: 2008 IEEE 24th International Conference on Data Engineering; Apr 7-12, 2008; Cancun, Mexico. [doi: 10.1109/icde.2008.4497459]
160. Hay M, Miklau G, Jensen D, Towsley D, Weis P. Resisting structural re-identification in anonymized social networks. Proc VLDB Endow 2008 Aug 14;1(1):102-114. [doi: 10.14778/1453856.1453873]
161. Feder T, Nabar S, Terzi E. Anonymizing graphs. arXiv. 2008. URL: https://arxiv.org/abs/0810.5578 [accessed 2021-08-30]
162. Stokes K, Torra V. Reidentification and k-anonymity: a model for disclosure risk in graphs. Soft Comput 2012 May 1;16(10):1657-1670. [doi: 10.1007/s00500-012-0850-4]
163. Bettini C, Wang X, Jajodia S. Protecting privacy against location-based personal identification. In: Secure Data Management. SDM 2005. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2005.
164. ARX: a comprehensive tool for anonymizing biomedical data. Institute of Medical Statistics and Epidemiology. 2015. URL: https://pdfs.semanticscholar.org/dbe3/a46ea167e70a086ce96a2ffed56be2114c0f.pdf [accessed 2021-08-30]
165. Emam KE, Arbuckle L. Anonymizing Health Data: Case Studies and Methods to Get You Started. Sebastopol, California, United States: O'Reilly Media; 2014.
166. Erkin Z, Veugen T, Toft T, Lagendijk RL. Generating private recommendations efficiently using homomorphic encryption and data packing. IEEE Trans Inform Forensic Secur 2012 Jun;7(3):1053-1066. [doi: 10.1109/tifs.2012.2190726]
167. Paillier P. Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques. 1999 Presented at: Annual International Conference on the Theory and Applications of Cryptographic Techniques; May 2-6, 1999; Prague, Czech Republic. [doi: 10.1007/3-540-48910-x_16]
168. Kung S. Compressive privacy: from information\/estimation theory to machine learning [lecture notes]. IEEE Signal Process Mag 2017 Jan;34(1):94-112. [doi: 10.1109/msp.2016.2616720]
169. Kung S, Chanyaswad T, Chang JM, Wu P. Collaborative PCA/DCA learning methods for compressive privacy. ACM Trans Embed Comput Syst 2017 Jul;16(3):1-18. [doi: 10.1145/2996460]
170. Pinto A. A comparison of anonymization protection principles. In: Proceedings of the 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI). 2012 Presented at: 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI); Aug 8-10, 2012; Las Vegas, NV, USA. [doi: 10.1109/iri.2012.6303012]
171. Elliot M, Domingo-Ferrer J. The future of statistical disclosure control. arXiv. 2018. URL: https://arxiv.org/abs/1812.09204 [accessed 2021-08-30]
172. Ayala-Rivera V, McDonagh P, Cerqueus T, Murphy L. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. Trans Data Privacy 2014 Dec;7(3):337-370. [doi: 10.5555/2870614.2870620]
173. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06). 2007 Presented at: 22nd International Conference on Data Engineering (ICDE'06); Apr 3-7, 2006; Atlanta, GA, USA. [doi: 10.1109/ICDE.2006.1]
174. Wong R, Li J, Fu A, Wang K. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006 Presented at: KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 20-23, 2006; Philadelphia PA USA. [doi: 10.1145/1150402.1150499]
175. Sweeney L. Computational disclosure control: a primer on data privacy protection. Thesis and Dissertations - Massachusetts Institute of Technology. 2001. URL: https://dspace.mit.edu/handle/1721.1/8589 [accessed 2021-08-30]
176. Xiao X, Tao Y. M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. 2007 Presented at: SIGMOD '07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data; Jun 11-14, 2007; Beijing China. [doi: 10.1145/1247480.1247556]
177. Zhang Q, Koudas N, Srivastava D, Yu T. Aggregate query answering on anonymized tables. In: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering. 2007 Presented at: 2007 IEEE 23rd International Conference on Data Engineering; Apr 15-20, 2007; Istanbul, Turkey. [doi: 10.1109/icde.2007.367857]
178. Qishan Z, Zhensi L, Qunhua Z, Hong L. (K, G)-anonymity model based on grey relational analysis. In: Proceedings of 2013 IEEE International Conference on Grey systems and Intelligent Services (GSIS). 2013 Presented at: Proceedings of 2013 IEEE International Conference on Grey systems and Intelligent Services (GSIS); Nov 15-17, 2013; Macao, China. [doi: 10.1109/gsis.2013.6714730]
179. Han J, Cen T, Yu H. Research in microaggregation algorithms for k-anonymization. Acta Electronica Sinica 2008;36(10):2021-2029 [FREE Full text]
180. Nergiz M, Clifton C, Nergiz A. Multirelational k-anonymity. IEEE Trans Knowl Data Eng 2009 Aug;21(8):1104-1117. [doi: 10.1109/tkde.2008.210]

181. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn K. ARX--a comprehensive tool for anonymizing biomedical data. AMIA Annu Symp Proc 2014 Nov 14;2014:984-993 [FREE Full text] [Medline: 25954407]
182. Introduction of Anonimatron - The free, extendable, open source data anonymization tool. GitHub. URL: https://realrolfje.github.io/anonimatron/ [accessed 2021-09-29]
183. Kearns M, Pai M, Roth A, Ullman J. Mechanism design in large games: incentives and privacy. In: Proceedings of the 5th conference on Innovations in Theoretical Computer Science. 2014 Presented at: ITCS '14: Proceedings of the 5th conference on Innovations in Theoretical Computer Science; Jan 12-14, 2014; Princeton New Jersey USA. [doi: 10.1145/2554797.2554834]
184. Kasperbauer TJ. Protecting health privacy even when privacy is lost. J Med Ethics 2020 Nov;46(11):768-772. [doi: 10.1136/medethics-2019-105880] [Medline: 31806677]
185. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020 Oct 22;63(11):139-144. [doi: 10.1145/3422622]
186. D'Acquisto G, Naldi M. A conceptual framework for assessing anonymization-utility trade-offs based on principal component analysis. arXiv. 2019. URL: https://arxiv.org/abs/1903.11700?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+arxiv%2FQSXk+%28ExcitingAds%21+cs+updates+on+arXiv.org%29 [accessed 2021-08-30]
187. Matatov N, Rokach L, Maimon O. Privacy-preserving data mining: a feature set partitioning approach. Inf Sci 2010 Jul 15;180(14):2696-2720. [doi: 10.1016/j.ins.2010.03.011]
188. Iyengar V. Transforming data to satisfy privacy constraints. In: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002 Presented at: KDD02: The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Jul 23-26, 2002; Edmonton Alberta Canada. [doi: 10.1145/775047.775089]
189. Nergiz ME, Clifton C. Thoughts on k-anonymization. Data Know Eng 2007 Dec;63(3):622-645. [doi: 10.1016/j.datak.2007.03.009]
190. Mahesh R, Meyyappan T. Anonymization technique through record elimination to preserve privacy of published data. In: Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering. 2013 Presented at: 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering; Feb 21-22, 2013; Salem, India. [doi: 10.1109/icprime.2013.6496495]
191. Bayardo R, Agrawal R. Data privacy through optimal k-anonymization. In: Proceedings of the 21st International Conference on Data Engineering (ICDE'05). 2005 Presented at: 21st International Conference on Data Engineering (ICDE'05); Apr 5-8, 2005; Tokyo, Japan. [doi: 10.1109/icde.2005.42]
192. Foygel R, Srebro N, Salakhutdinov R. Matrix reconstruction with the local max norm. arXiv. 2012. URL: https://arxiv.org/abs/1210.5196 [accessed 2021-08-30]
193. Zwillinger D. CRC Standard Mathematical Tables and Formulas. Boca Raton: CRC Press; Jan 2018.
194. Kshirsagar AM. Correlation between two vector variables. J R Stat Soc Series B Stat Methodol 2018 Dec 05;31(3):477-485. [doi: 10.1111/j.2517-6161.1969.tb00807.x]
195. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951 Mar;22(1):79-86. [doi: 10.1214/aoms/1177729694]
196. Duchi J. Derivations for linear algebra and optimization. Stanford. 2007. URL: http://ai.stanford.edu/~jduchi/projects/general_notes.pdf [accessed 2021-08-30]
197. Hu X, Fu C, Zhu L, Heng PA. Depth-attentional features for single-image rain removal. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019 Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019; Long Beach, CA, USA. [doi: 10.1109/cvpr.2019.00821]
198. Yang W, Tan RT, Feng J, Guo Z, Yan S, Liu J. Joint rain detection and removal from a single image with contextualized deep networks. IEEE Trans Pattern Anal Mach Intell 2020 Jun 1;42(6):1377-1393. [doi: 10.1109/tpami.2019.2895793]
199. Wang T, Yang X, Xu K, Chen S, Zhang Q, Lau R. Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019 Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019; Long Beach, CA, USA. [doi: 10.1109/cvpr.2019.01255]
200. Fu X, Liang B, Huang Y, Ding X, Paisley J. Lightweight pyramid networks for image deraining. IEEE Trans Neural Netw Learn Syst 2020 Jun;31(6):1794-1807. [doi: 10.1109/tnnls.2019.2926481]
201. Kim J, Kwon LJ, Mu LK. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, USA. [doi: 10.1109/cvpr.2016.182]
202. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y. Image super-resolution using very deep residual channel attention networks. In: Computer Vision – ECCV 2018. Cham: Springer; 2018.
203. Shen H, Sun S, Wang J, Dong F, Lei B. Comparison of image quality objective evaluation. In: Proceedings of the 2009 International Conference on Computational Intelligence and Software Engineering. 2009 Presented at: 2009 International

Conference on Computational Intelligence and Software Engineering; Dec 11-13, 2009; Wuhan, China. [doi: 10.1109/cise.2009.5366163]
204. Wang B, Wang Z, Liao Y, Lin X. HVS-based structural similarity for image quality assessment. In: Proceedings of the 2008 9th International Conference on Signal Processing. 2008 Presented at: 2008 9th International Conference on Signal Processing; Oct 26-29, 2008; Beijing, China. [doi: 10.1109/icosp.2008.4697344]
205. Winkler S. Issues in vision modeling for perceptual video quality assessment. Signal Process 1999 Oct;78(2):231-252 [FREE Full text] [doi: 10.1016/s0165-1684(99)00062-6]
206. Zhang K, Gool L, Timofte R. Deep unfolding network for image super-resolution. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020; Seattle, WA, USA. [doi: 10.1109/cvpr42600.2020.00328]
207. Wang Z, Bovik A, Sheikh H, Simoncelli E. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004 Apr;13(4):600-612. [doi: 10.1109/tip.2003.819861] [Medline: 15376593]
208. Yang W, Tan RT, Wang S, Fang Y, Liu J. Single image deraining: from model-based to data-driven and beyond. IEEE Trans Pattern Anal Mach Intell 2020 May 19:1. [doi: 10.1109/tpami.2020.2995190]
209. P.800 : Methods for subjective determination of transmission quality. International Telecommunication Union. 1996. URL: https://www.itu.int/rec/T-REC-P.800-199608-I [accessed 2021-09-30]
210. RECOMMENDATION ITU-R BT.500-11: Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union. 2002. URL: http://www.gpds.ene.unb.br/databases/2012-UNB-Varium-Exp/Exp3-Delft/00-report-alexandre/Papers---Judith/Subjective%20Studies/ITU-Recommendation---BT500-11.pdf [accessed 2021-09-30]
211. Yang W, Yuan Y, Ren W, Liu J, Scheirer WJ, Wang Z, et al. Advancing image understanding in poor visibility environments: a collective benchmark study. IEEE Trans Image Process 2020 Mar 27;29:5737-5752. [doi: 10.1109/TIP.2020.2981922] [Medline: 32224457]
212. Kaufman A, Fattal R. Deblurring using analysis-synthesis networks pair. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 13-19, 2020; Seattle, WA, USA. [doi: 10.1109/CVPR42600.2020.00585]
213. Yang R, Mentzer F, Gool LV, Timofte R. Learning for video compression with hierarchical quality and recurrent enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 13-19, 2020; Seattle, WA, USA. [doi: 10.1109/CVPR42600.2020.00666]
214. Chaudhuri K, Sarwate A, Sinha K. A near-optimal algorithm for differentially-private principal components. J Mach Learn Res 2013;14(1):2905-2943. [doi: 10.4016/38611.01]
215. Dwork C, Talwar K, Thakurta A, Zhang L. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In: Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing. 2014 Presented at: STOC '14: Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing; May 31- Jun 3, 2014; New York. [doi: 10.1145/2591796.2591883]
216. Wei L, Sarwate A, Corander J, Hero A, Tarokh V. Analysis of a privacy-preserving PCA algorithm using random matrix theory. In: Proceedings of the 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP). 2016 Presented at: 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP); Dec 7-9, 2016; Washington, DC, USA. [doi: 10.1109/globalsip.2016.7906058]
217. Al-Rubaie M, Wu P, Chang J, Kung SY. Privacy-preserving PCA on horizontally-partitioned data. In: Proceedings of the 2017 IEEE Conference on Dependable and Secure Computing. 2017 Presented at: 2017 IEEE Conference on Dependable and Secure Computing; Aug 7-10, 2017; Taipei, Taiwan. [doi: 10.1109/desec.2017.8073817]
218. Grammenos A, Mendoza-Smith R, Mascolo C, Crowcroft J. Federated PCA with adaptive rank estimation. CoRR. 2019. URL: https://www.researchgate.net/publication/334558717_Federated_PCA_with_Adaptive_Rank_Estimation [accessed 2021-09-29]
219. Liu Y, Chen C, Zheng L, Wang L, Zhou J, Liu G. Privacy preserving PCA for multiparty modeling. arXiv. 2020. URL: https://arxiv.org/abs/2002.02091 [accessed 2021-08-30]
220. Chen C, Liu Z, Zhao P, Zhou J, Li X. Privacy preserving point-of-interest recommendation using decentralized matrix factorization. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. 2018 Presented at: Thirty-Second AAAI Conference on Artificial Intelligence; Feb 2-7, 2018; New Orleans, Louisiana, USA URL: https://ojs.aaai.org/index.php/AAAI/article/view/11244
221. Mohassel P, Zhang Y. SecureML: a system for scalable privacy-preserving machine learning. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). 2017 Presented at: 2017 IEEE Symposium on Security and Privacy (SP); May 22-26, 2017; San Jose, CA, USA. [doi: 10.1109/sp.2017.12]
222. Xing K, Hu C, Yu J, Cheng X, Zhang F. Mutual privacy preserving k -means clustering in social participatory sensing. IEEE Trans Ind Inf 2017 Aug 18;13(4):2066-2076. [doi: 10.1109/tii.2017.2695487]
223. Li P, Chen Z, Yang LT, Zhao L, Zhang Q. A privacy-preserving high-order neuro-fuzzy c-means algorithm with cloud computing. Neurocomputing 2017 Sep;256:82-89. [doi: 10.1016/j.neucom.2016.08.135]

224. Manikandan V, Porkodi V, Mohammed A, Sivaram M. Privacy preserving data mining using threshold based fuzzy cmeans clustering. ICTACT J Soft Comput 2018 Oct;9(1):0253. [doi: 10.21917/ijsc.2018.0253]
225. Anuradha P, Srinivas Y, Prasad MK. A frame work for preserving privacy in social media using generalized gaussian mixture model. Int J Adv Comput Sci Appl 2015;6(7):68-71. [doi: 10.14569/ijacsa.2015.060711]
226. Hahn S, Lee J. GRAFFL: gradient-free federated learning of a Bayesian generative model. arXiv. 2020. URL: https://arxiv.org/abs/2008.12925 [accessed 2021-08-30]
227. Ahmed F, Liu A, Jin R. Publishing social network graph eigenspectrum with privacy guarantees. IEEE Trans Netw Sci Eng 2020;7(2):892-906. [doi: 10.1109/tnse.2019.2901716]
228. Li J, Wei J, Ye M, Liu W, Hu X. Privacy‐preserving constrained spectral clustering algorithm for large‐scale data sets. IET Inf Secur 2020 May;14(3):321-331. [doi: 10.1049/iet-ifs.2019.0255]
229. Li Y, Ma J, Miao Y, Wang Y, Liu X, Choo KR. Similarity search for encrypted images in secure cloud computing. IEEE Trans Cloud Comput 2020 Apr 27:1. [doi: 10.1109/tcc.2020.2989923]
230. Almutairi N, Coenen F, Dures K. Data clustering using homomorphic encryption and secure chain distance matrices. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR. 2018 Presented at: 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR; Sep 18-20, 2018; Seville, Spain. [doi: 10.5220/0006890800410050]
231. Almutairi N, Coenen F, Dures K. A cryptographic ensemble for secure third party data analysis: collaborative data clustering without data owner participation. Data Knowl Eng 2020 Mar;126:101734. [doi: 10.1016/j.datak.2019.101734]
232. Zhu Y, Yu X, Chandraker M, Wang YX. Private-kNN: practical differential privacy for computer vision. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020; Seattle, WA, USA. [doi: 10.1109/cvpr42600.2020.01187]
233. Bian S, Wang T, Hiromoto M, Shi Y, Sato T. ENSEI: efficient secure inference via frequency-domain homomorphic convolution for privacy-preserving visual recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020; Seattle, WA, USA. [doi: 10.1109/cvpr42600.2020.00942]
234. Malhotra A, Chhabra S, Vatsa M, Singh R. On privacy preserving anonymization of finger-selfies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 14-19, 2020; Seattle, WA, USA. [doi: 10.1109/cvprw50498.2020.00021]
235. mp2893 / medgan. GitHub. URL: https://github.com/mp2893/medgan [accessed 2021-08-30]
236. Baowaly M, Lin CC, Liu CL, Chen KT. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc 2019 Mar 01;26(3):228-241 [FREE Full text] [doi: 10.1093/jamia/ocy142] [Medline: 30535151]
237. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein GANs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17: 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach California USA.
238. Shin H, Tenenholtz N, Rogers J, Schwarz C, Senjem M, Gunter J. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: Simulation and Synthesis in Medical Imaging. Cham: Springer; 2018.
239. Yoon J, Drumright LN, van der Schaar M. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). IEEE J Biomed Health Inform 2020 Aug;24(8):2378-2388. [doi: 10.1109/JBHI.2020.2980262] [Medline: 32167919]
240. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes 2019 Jul;12(7):e005122 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.118.005122] [Medline: 31284738]
241. Chin-Cheong K, Sutter T, Vogt J. Generation of differentially private heterogeneous electronic health records. arXiv. 2020. URL: https://tinyurl.com/j825avzj [accessed 2021-08-30]
242. Jordon J, Yoon J, van der Schaar M. PATE-GAN: generating synthetic data with differential privacy guarantees. In: Proceedings of the International Conference on Learning Representations (2019). 2019 Presented at: International Conference on Learning Representations (2019); May 6-9, 2019; New Orleans, Louisiana, USA URL: https://openreview.net/forum?id=S1zk9iRqF7
243. Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially private generative adversarial network. arXiv. 2018. URL: https://arxiv.org/abs/1802.06739 [accessed 2021-08-30]
244. Fan L. A survey of differentially private generative adversarial networks. In: Proceedings of the 2020 AAAI Workshop on Privacy-Preserving Artificial Intelligence. 2020 Presented at: AAAI Workshop on Privacy-Preserving Artificial Intelligence; Feb. 7-12, 2020; New York USA URL: https://www2.isye.gatech.edu/~fferdinando3/cfp/PPAI20/papers/paper_9.pdf
245. Tseng B, Wu P. Compressive privacy generative adversarial network. IEEE Trans Inform Forensic Secur 2020 Jan 20;15:2499-2513. [doi: 10.1109/tifs.2020.2968188]

246. Maximov M, Elezi I, Leal-Taixe L. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020 Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020; Seattle, WA, USA. [doi: 10.1109/cvpr42600.2020.00549]
247. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv. 2014. URL: https://arxiv.org/abs/1411.1784 [accessed 2021-08-30]
248. Domingo-Ferrer J. Microaggregation for database and location privacy. In: Next Generation Information Technologies and Systems. Berlin, Heidelberg: Springer; 2006.
249. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969 Dec;64(328):1183-1210. [doi: 10.1080/01621459.1969.10501049]
250. Domingo-Ferrer J. A survey of inference control methods for privacy-preserving data mining. In: Privacy-Preserving Data Mining. Advances in Database Systems. Boston, MA: Springer; 2008.
251. Skinner C, Holmes D. Estimating the re-identification risk per record in microdata. J Off Stat 1998;14(4):361-372.
252. Documentation. Amnesia. URL: https://amnesia.openaire.eu/about-documentation.html [accessed 2021-08-30]
253. Anonimatron. GitHub. URL: https://realrolfje.github.io/anonimatron/ [accessed 2021-08-30]
254. Prasser F, Kohlmayer F. Putting statistical disclosure control into practice: the ARX data anonymization tool. In: Medical Data Privacy Handbook. Cham: Springer; 2015.
255. Dankar FK, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak 2012 Jul 09;12:66 [FREE Full text] [doi: 10.1186/1472-6947-12-66] [Medline: 22776564]
256. arx-deidentifier / arx. GitHub. URL: https://github.com/arx-deidentifier/arx [accessed 2021-08-30]
257. Sánchez D, Martínez S, Domingo-Ferrer J, Soria-Comas J, Batet M. µ-ANT: semantic microaggregation-based anonymization tool. Bioinformatics 2020 Mar 01;36(5):1652-1653. [doi: 10.1093/bioinformatics/btz792] [Medline: 31621826]
258. CrisesUrv / microaggregation-based_anonymization_tool. GitHub. URL: https://github.com/CrisesUrv/microaggregation-based_anonymization_tool [accessed 2021-08-30]
259. Package 'sdcMicro'. CRAN. 2021. URL: https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf [accessed 2021-08-30]
260. µ-ARGUS home page. Statistical Disclosure Control. 2018. URL: https://research.cbs.nl/casc/mu.htm [accessed 2021-08-30]
261. UT Dallas anonymization toolbox. UT Dallas. URL: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/anonManual.pdf [accessed 2021-08-30]
262. Dai C, Ghinita G, Bertino E, Byun J, Li N. TIAMAT: a tool for interactive analysis of microdata anonymization techniques. Proc VLDB Endow 2009 Aug;2(2):1618-1621. [doi: 10.14778/1687553.1687607]
263. Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA. Flexible data anonymization using ARX—current status and challenges ahead. Softw Pract Exper 2020 Feb 25;50(7):1277-1304. [doi: 10.1002/spe.2812]
264. Poulis G, Gkoulalas-Divanis A, Loukides G, Skiadopoulos S, Tryfonopoulos C. SECRETA: a system for evaluating and comparing relational and transaction anonymization algorithms. In: Medical Data Privacy Handbook. Cham: Springer; 2015.
265. Gkoulalas-Divanis A, Braghin S. IPV: a system for identifying privacy vulnerabilities in datasets. IBM J Res Dev 2016 Jul 27;60(4):1-10. [doi: 10.1147/jrd.2016.2576818]
266. Francis P, Probst ES, Munz R. Diffix: high-utility database anonymization. In: Privacy Technologies and Policy. Cham: Springer; 2017.
267. Francis P, Probst-Eide S, Obrok P, Berneanu C, Juric S, Munz R. Diffix-Birch: extending Diffix-Aspen. arXiv. 2018. URL: https://arxiv.org/abs/1806.02075 [accessed 2021-08-30]
268. Aircloak home page. Aircloak. URL: https://aircloak.com/ [accessed 2021-08-30]
269. NLM-Scrubber downloads. NLM-Scrubber. URL: https://scrubber.nlm.nih.gov/files/ [accessed 2021-08-30]
270. OpenPseudonymiser home page. OpenPseudonymiser. URL: https://www.openpseudonymiser.org/About.aspx [accessed 2021-08-30]
271. Vardalachakis M, Kondylakis H, Koumakis L, Kouroubali A, Katehakis D. ShinyAnonymizer: a tool for anonymizing health data. In: Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE. 2019 Presented at: 5th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE,; May 2-4, 2019; Heraklion, Crete, Greece. [doi: 10.5220/0007798603250332]
272. Karwatka P. GDPR quick wins for software developers and teams. LinkedIn. 2017. URL: https://www.linkedin.com/pulse/gdpr-quick-wins-software-developers-teams-piotr-karwatka/ [accessed 2021-08-30]
273. A way to exclude rows matching certain criteria? GitHub. 2018. URL: https://github.com/DivanteLtd/anonymizer/issues/4 [accessed 2021-08-30]
274. Xiao X, Wang G, Gehrke J. Interactive anonymization of sensitive data. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. 2009 Presented at: SIGMOD '09: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data; Jun 29-Jul 2, 2009; Providence Rhode Island USA. [doi: 10.1145/1559845.1559979]

275. Cornell anonymization toolkit. Source Forge. URL: https://sourceforge.net/projects/anony-toolkit/files/Documents/ [accessed 2021-08-30]
276. Yao A. Protocols for secure computations. In: Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science. 1982 Presented at: 23rd IEEE Symposium on Foundations of Computer Science; Nov 3-5, 1982; San Francisco, CA, USA. [doi: 10.1109/sfcs.1982.38]
277. Keller M, Pastro V, Rotaru D. Overdrive: making SPDZ great again. In: Advances in Cryptology – EUROCRYPT 2018. Cham: Springer; 2018.
278. Ryffel T, Trask A, Dahl M, Wagner B, Mancuso J, Rueckert D. A generic framework for privacy preserving deep learning. CoRR 2018:4017.
279. Crypten: a research tool for secure machine learning in PyTorc. Crypten. URL: https://crypten.ai [accessed 2021-08-30]
280. Dahl M, Mancuso J, Dupis Y, Decoste B, Giraud M, Livingstone I, et al. Private machine learning in TensorFlow using secure computation. Privacy Preserving Machine Learning Workshop. 2018. URL: https://ppml-workshop.github.io/ppml18/slides/56.pdf [accessed 2021-08-30]
281. Kumar N, Rathee M, Chandran N, Gupta D, Rastogi A, Sharma R. CrypTFlow: secure TensorFlow inference. In: Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP). 2020 Presented at: 2020 IEEE Symposium on Security and Privacy (SP); May 18-21, 2020; San Francisco, CA, USA. [doi: 10.1109/sp40000.2020.00092]
282. Zhu X, Iordanescu G, Karmanov I, Zawaideh M. Using Microsoft AI to build a lung-disease prediction model using chest X-Ray images. Microsoft. 2018. URL: https://docs.microsoft.com/en-us/archive/blogs/machinelearning/using-microsoft-ai-to-build-a-lung-disease-prediction-model-using-chest-x-ray-images [accessed 2021-08-31]
283. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
284. Hong C, Huang Z, Lu W, Qu H, Ma L, Dahl M. Privacy-preserving collaborative machine learning on genomic data using TensorFlow. In: Proceedings of the ACM Turing Celebration Conference - China. 2020 Presented at: ACM TURC'20: ACM Turing Celebration Conference - China; May 22-24, 2020; Hefei China. [doi: 10.1145/3393527.3393535]
285. Dowlin N, Gilad-Bachrach R, Laine K, Lauter K, Naehrig M, Wernsing J. CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. In: Proceedings of the International Conference on Machine Learning. 2016 Presented at: International Conference on Machine Learning; June 20-22, 2016; New York, USA p. 201-210 URL: https://proceedings.mlr.press/v48/gilad-bachrach16.html
286. Hesamifard E, Takabi H, Ghasemi M, Wright R. Privacy-preserving machine learning as a service. Proc Priv Enhanc Technol 2018 Apr 28(3):123-142. [doi: 10.1515/popets-2018-0024]
287. Dathathri R, Saarikivi O, Chen H, Laine K, Lauter K, Maleki S, et al. CHET: an optimizing compiler for fully-homomorphic neural-network inferencing. In: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation. 2019 Jun Presented at: PLDI 2019: 40th ACM SIGPLAN Conference on Programming Language Design and Implementation; Jun 22-26, 2019; Phoenix AZ USA. [doi: 10.1145/3314221.3314628]
288. microsoft / SEAL. GitHub. 2019. URL: https://github.com/Microsoft/SEAL [accessed 2021-08-30]
289. Cheon J, Kim A, Kim M, Song Y. Homomorphic encryption for arithmetic of approximate numbers. In: Advances in Cryptology – ASIACRYPT 2017. Cham: Springer; 2017.
290. What is encryption? Information Commissioner's Office. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/encryption/what-is-encryption/#%233 [accessed 2021-08-30]
291. Art. 5 GDPR Principles relating to processing of personal data. Intersoft Consulting. URL: https://gdpr-info.eu/art-5-gdpr/ [accessed 2021-08-30]
292. Data Protection Act 2018. UK Government Legislation. 2018. URL: https://www.legislation.gov.uk/ukpga/2018/12/pdfs/ukpga_20180012_en.pdf [accessed 2021-08-30]
293. Anonymisation standard for publishing health and social care data. NHS Digital. 2013. URL: https://digital.nhs.uk/binaries/content/assets/legacy/pdf/b/o/1523202010spec.pdf [accessed 2021-08-30]
294. Confidentiality and disclosure of information policy. National Health Service. URL: https://www.newhamccg.nhs.uk/Downloads/News-and-Publications/Policies-and-procedures/Confidentiality%20and%20Disclosure%20of%20Information%20Policy.pdf [accessed 2021-08-30]
295. Pseudonymisation and anonymisation of data - procedure. Tavistock and Portman NHS Foundation Trust. 2019. URL: https://tavistockandportman.nhs.uk/documents/1301/pseudomymisation-anonymisation-data-procedure-Jan-19.pdf [accessed 2021-08-30]
296. Pseudonymisation policy. Hounslow NHS Clinical Commissioning Group. 2021. URL: https://www.hounslowccg.nhs.uk/media/164202/Pseudonymisation-and-Anonymisation-Policy-v11-Final-05022021.pdf [accessed 2021-08-30]
297. Guidance for using patient data. NHS Health Research Authority. 2021. URL: https://www.hra.nhs.uk/covid-19-research/guidance-using-patient-data/ [accessed 2021-08-30]
298. Data protection and coronavirus – advice for organisations. Information Commissioner's Office. 2021. URL: https://ico.org.uk/global/data-protection-and-coronavirus-information-hub/coronavirus-recovery-data-protection-advice-for-organisations/ [accessed 2021-08-30]

299. Statement by the EDPB Chair on the processing of personal data in the context of the COVID-19 outbreak. European Data Protection Board. 2020. URL: https://edpb.europa.eu/news/news/2020/statement-edpb-chair-processing-personal-data-context-covid-19-outbreak_en [accessed 2021-08-30]
300. da Silva JE, de Sá JP, Jossinet J. Classification of breast tissue by electrical impedance spectroscopy. Med Biol Eng Comput 2000 Jan;38(1):26-30. [doi: 10.1007/bf02344684]
301. Antal B, Hajdu A. An ensemble-based system for automatic screening of diabetic retinopathy. Knowl Based Syst 2014 Apr;60:20-27. [doi: 10.1016/j.knosys.2013.12.023]
302. Zuo Z, Li J, Anderson P, Yang L, Naik N. Grooming detection using fuzzy-rough feature selection and text classification. In: Proceedings of the 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2018 Presented at: 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); Jul 8-13, 2018; Rio de Janeiro, Brazil. [doi: 10.1109/fuzz-ieee.2018.8491591]
303. Zuo Z, Li J, Wei B, Yang L, Chao F, Naik N. Adaptive activation function generation for artificial neural networks through fuzzy inference with application in grooming text categorisation. In: Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2019 Presented at: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); Jun 23-26, 2019; New Orleans, LA, USA. [doi: 10.1109/fuzz-ieee.2019.8858838]
304. Domingo-Ferrer J, Torra V. A quantitative comparison of disclosure control methods for microdata. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam, Netherlands: Elsevier; 2001:111-134.
305. Duncan GT, Fienberg SE, Krishnan R, Padman R, Roehrig SF. Disclosure limitation methods and information loss for tabular data. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: Elsevier; 2001:135-166.
306. Duncan GT, Stokes SL. Disclosure risk vs. data utility: the R-U confidentiality map as applied to topcoding. Chance 2012 Sep 20;17(3):16-20. [doi: 10.1080/09332480.2004.10554908]
307. Harder F, Bauer M, Park M. Interpretable and differentially private predictions. arXiv. 2020. URL: https://arxiv.org/abs/1906.02004 [accessed 2021-08-30]
308. Artificial intelligence: how to get it right. National Health Service. 2019 Jan. URL: https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/understand-ai/artificial-intelligence-how-get-it-right/ [accessed 2021-08-30]
309. GDPR right to be informed. Intersoft Consulting. URL: https://gdpr-info.eu/issues/right-to-be-informed/ [accessed 2021-08-30]

Abbreviations

CP: compressive privacy
DP: differential privacy
DPA: Data Protection Act
EHR: electronic health record
FHE: fully homomorphic encryption
GAN: generative adversarial network
GDPR: General Data Protection Regulation
GP: general practitioner
HE: homomorphic encryption
HIPAA: Health Information Portability and Accountability Act
ICO: Information Commissioner's Office
medGAN: medical generative adversarial network
NHS: National Health Service
PCA: principal component analysis
PRAM: postrandomization method
QI: quasi-identifier
RQ: review question
SECRETA: System for Evaluating and Comparing RElational and Transaction Anonymization
SLM: systematic literature mapping
SLR: systematic literature review
SMPC: secure multiparty computation
TF: TensorFlow
μ-ANT: microaggregation-based anonymization tool

Edited by G Eysenbach; submitted 23.04.21; peer-reviewed by I Schiering; comments to author 14.05.21; revised version received 21.06.21; accepted 02.08.21; published 15.10.21.

Please cite as:
Zuo Z, Watson M, Budgen D, Hall R, Kennelly C, Al Moubayed N
Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study
JMIR Med Inform 2021;9(10):e29871
URL: https://medinform.jmir.org/2021/10/e29871
doi:10.2196/29871
PMID:34652278

Original Paper

An Automated Line-of-Therapy Algorithm for Adults With Metastatic Non–Small Cell Lung Cancer: Validation Study Using Blinded Manual Chart Review

Weilin Meng[1], MEng; Kelly M Mosesso[2], MA; Kathleen A Lane[2], MS; Anna R Roberts[3], MIS; Ashley Griffith[3], MHA; Wanmei Ou[1], PhD; Paul R Dexter[3,4,5], MD

[1]Center for Observational and Real-World Evidence, Merck & Co, Inc, Kenilworth, NJ, United States

[2]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, United States

[3]Regenstrief Institute, Inc, Indianapolis, IN, United States

[4]Eskenazi Health, Indianapolis, IN, United States

[5]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, United States

Corresponding Author:
Paul R Dexter, MD
Regenstrief Institute, Inc
1101 West 10th Street
Indianapolis, IN, 46202-4800
United States
Phone: 1 317 274 9000
Email: prdexter@regenstrief.org

Abstract

Background: Extraction of line-of-therapy (LOT) information from electronic health record and claims data is essential for determining longitudinal changes in systemic anticancer therapy in real-world clinical settings.

Objective: The aim of this retrospective cohort analysis is to validate and refine our previously described open-source LOT algorithm by comparing the output of the algorithm with results obtained through blinded manual chart review.

Methods: We used structured electronic health record data and clinical documents to identify 500 adult patients treated for metastatic non–small cell lung cancer with systemic anticancer therapy from 2011 to mid-2018; we assigned patients to training (n=350) and test (n=150) cohorts, randomly divided proportional to the overall ratio of simple:complex cases (n=254:246). Simple cases were patients who received one LOT and no maintenance therapy; complex cases were patients who received more than one LOT and/or maintenance therapy. Algorithmic changes were performed using the training cohort data, after which the refined algorithm was evaluated against the test cohort.

Results: For simple cases, 16 instances of discordance between the LOT algorithm and chart review prerefinement were reduced to 8 instances postrefinement; in the test cohort, there was no discordance between algorithm and chart review. For complex cases, algorithm refinement reduced the discordance from 68 to 62 instances, with 37 instances in the test cohort. The percentage agreement between LOT algorithm output and chart review for patients who received one LOT was 89% prerefinement, 93% postrefinement, and 93% for the test cohort, whereas the likelihood of precise matching between algorithm output and chart review decreased with an increasing number of unique regimens. Several areas of discordance that arose from differing definitions of LOTs and maintenance therapy could not be objectively resolved because of a lack of precise definitions in the medical literature.

Conclusions: Our findings identify common sources of discordance between the LOT algorithm and clinician documentation, providing the possibility of targeted algorithm refinement.

(JMIR Med Inform 2021;9(10):e29017) doi:10.2196/29017

KEYWORDS

automated algorithm; line of therapy; longitudinal changes; manual chart review; non–small cell lung cancer; systemic anticancer therapy

Introduction

Lung cancer is the most common cause of cancer-related deaths worldwide [1], accounting for almost 2 million deaths annually [2,3]. Non–small cell lung cancer (NSCLC) represents approximately 85% of all lung cancer cases [4]. Treatment for advanced NSCLC is increasingly based on molecular patterns, including therapies that target mutations such as *EGFR* and *ALK* genomic aberrations, as well as inhibitors of the programmed death 1 (PD-1) pathway, particularly for patients whose tumors have high levels of PD-ligand 1 expression [5]. Although survival for patients with advanced disease has improved, the need for continued therapeutic advances and research remains acute [4].

Cancer therapy is commonly classified into lines of therapy (LOTs), each comprising one or more cycles of a single agent or a combination systemic anticancer therapy (SACT) [6-8]. Extraction of LOT data from real-world transactional claims and electronic health records (EHRs) is essential for determining longitudinal SACT changes in real-world clinical care settings, but it is challenging because LOT information is often not clearly marked in structured data sets and therefore must be interpreted through clinical notes [6,9]. Researchers and clinicians use LOT information gathered retrospectively to determine the effectiveness of SACT regimens, identify trends in clinical practice patterns, identify eligible candidates for cancer trials, and conduct quality assurance to help ensure that patients receive optimal SACT [6,10,11]. Manual determination of LOT information for large numbers of patients is time consuming and often not feasible, prompting our own and others' searches for automated LOT algorithmic methods [6,12-15].

The objective of this study is to validate and refine our previously described open-source LOT algorithm [6] by comparing the LOT algorithm output with results obtained through independent blinded manual chart review.

Methods

Study Design, Patient Selection, and Data Extraction

After receiving approval from the Indiana University Institutional Review Board, we conducted a retrospective cohort analysis using structured EHR data and clinical documents from the Indiana Network for Patient Care, one of the largest and oldest health information exchanges in the United States [16-18]. The Indiana Network for Patient Care holds more than 13 billion data elements from more than 100 separate health care entities, including more than 130 million clinical documents providing data on nearly 15 million patients.

To validate the LOT algorithm, we identified adult patients treated for metastatic NSCLC with SACT and excluded patients who had received any SACT commonly used for small cell lung cancer, as described in Multimedia Appendix 1. To select the study cohort, we used the first iteration of the LOT algorithm to identify all the complex cases in the initial eligible population because we wanted to oversample patients with complex treatment sequences to train the algorithm. *Complex cases* were defined as patients who had either a maintenance therapy or more than one LOT, whereas *simple cases* were defined as those with only one LOT and no maintenance therapy. The complex cases were automatically selected for confirmation by chart review, and then simple cases were randomly chosen to complete the sample of 500 patients. The final determination of simple versus complex cases was thus made via chart review conducted by a physician (PRD).

Next, we extracted structured data commonly found in claims data and required by the LOT algorithm, including patient identifiers, SACT medications, and associated dates. For SACT medications, we filtered the SACT drug list to those used to treat metastatic NSCLC. The index date of the first-line (L1) treatment in this study corresponded to the date of initial SACT on or after recorded evidence of metastatic disease. For chart review purposes, we extracted all available clinical notes after the metastatic diagnosis date. In preparation for manual chart review, these clinical notes were loaded into nDepth, the Regenstrief natural language processing platform. This platform provides an efficient means of reviewing documents and capturing related information on a per-patient basis.

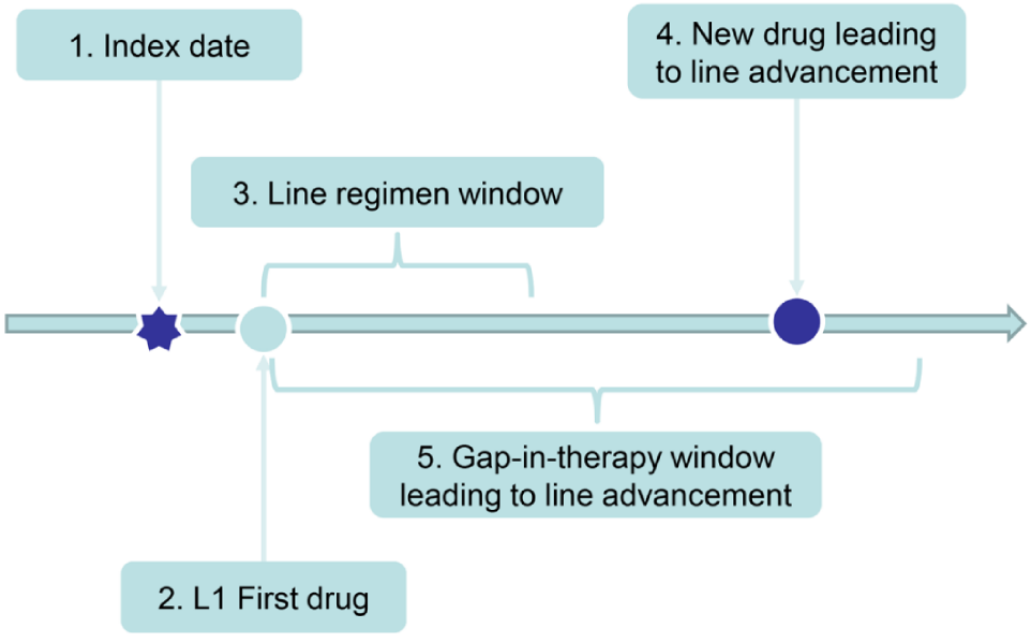
We then created a CSV file with patient identifier number, administration start date, administration end date (for oral drugs), and generic drug name as the column fields. This format is the minimum information required for the LOT algorithm input. Finally, we divided patients into a training cohort of 70% (350/500) patients and a test cohort of 30% (150/500) patients, using stratified sampling to keep the ratio of simple:complex cases the same in both cohorts. All algorithmic changes were performed using the training cohort data, after which the final version of the algorithm was evaluated against the test cohort.

LOT Algorithm

Investigators at Merck Sharp & Dohme Corp have internally developed automated business rules to identify LOT numbers, treatment regimens, and maintenance treatment for patients with cancer [6]. These rules consist of tumor-agnostic algorithmic processes that extract LOT information from claims databases or EHRs. Implemented as R (The R Foundation for Statistical Computing) and Python (Python Software Foundation) code routines, the LOT algorithm uses a modular design to facilitate plug-and-play tumor-specific customization. The LOT algorithm along with tumor-specific customizations is available as open-source software through GitHub [19].

An overview of the LOT algorithm can be roughly understood by breaking it down into the five basic modules depicted in Figure 1: (1) the index date is defined as the date of the metastatic NSCLC diagnosis; (2) the L1 first drug is defined as the first SACT drug claim recorded at any time on or after the index date; (3) the line regimen window is the time starting with the L1 first drug and extending forward in time, typically for 28 days, to capture any other drugs administered in combination with the first drug, and the resulting set of drugs defines that LOT treatment regimen; (4) line advancement occurs if a new drug not belonging to the treatment regimen is introduced; and (5) line advancement also occurs if a drug is administered after a long gap in therapy.

Figure 1. Schematic depicting the five basic modules of a line-of-therapy (LOT) algorithm. L1: first-line therapy. Reprinted with permission from Meng et al [6].



Within these modules, several parameters are available in the code that allow the adjustment and introduction of special cases and exceptions for the rules. Common adjustments relate to the detection of maintenance therapy, checking for drug switches early in a LOT, and adding exceptions to line advancement for gaps in therapy or when certain drug classes are added or substituted in a treatment regimen.

Blinded Manual Chart Review and Initial (Prerefinement) Validation of NSCLC Output

Blinded to results generated from the LOT algorithm, a physician (PRD) used the nDepth chart review functionality to review clinical notes for patients with metastatic NSCLC. The reviewer also had access to a spreadsheet that included the individual SACT medication names and dates of administration for each patient. The majority of detailed SACT LOT and maintenance therapy descriptions came from outpatient oncology notes. The reviewer extracted the following clinical information for each patient: (1) the sequence of SACT LOT and (2) maintenance therapy. He then formatted this information in a spreadsheet format identical to the LOT algorithm output to facilitate automated comparison.

For the initial (prerefinement) validation, we customized the NSCLC LOT algorithm parameters using the previously published criteria [6]. We then evaluated the output of the NSCLC LOT algorithm and compared it with the findings from chart review.

NSCLC LOT Algorithm Refinement and Subsequent (Postrefinement) Validation

After completion of the blinded, automated initial comparison between algorithm output and chart review for the patients in the training cohort (n=350), we identified issues accounting for any discordance between algorithm output and chart review. To evaluate the areas of discordance, we separated the cases into simple and complex categories. For each issue, we then refined the LOT algorithm using close review of the initial comparison results, iterative rerunning of the refined LOT algorithm against the original chart review results, discussion with internal experts, and targeted medical literature review. Researchers from Merck Sharp & Dohme, Indiana University, and Regenstrief reviewed the deidentified raw SACT data and arbitrated the differences between algorithm output and chart review through a series of meetings.

Statistical Analysis

Descriptive statistics were calculated for demographic and LOT characteristics, including means, SDs, ranges for continuous variables, and counts and percentages for categorical variables. One-way analysis of variance (ANOVA) and the Fisher exact test, as appropriate, were used to compare demographic characteristics and LOT counts between the training and test cohorts.

Intraclass correlation coefficients (ICCs) and the corresponding 95% CIs for the number of LOTs for each case based on the LOT algorithm and chart review were calculated. Percentage agreement and 95% CIs were calculated to compare the results from the LOT algorithm with the chart review. Agreement was defined as an exact match between the LOT algorithm output and physician chart review in terms of LOT number, regimen name, and maintenance therapy classification. Each LOT comprised the treatment as well as any subsequent maintenance therapy regimen.

Results

Cohort Selection

We identified 11,223 patients with at least one diagnosis code for lung cancer during the study period. Of these, 1461 patients had metastatic lung cancer as defined by diagnosis codes, metastatic criteria, and receipt of SACT 14 days before or any time after the index date. Of these 1461 patients, 897 patients also had NSCLC mentioned in unstructured patient notes.

To construct our final sample, the first iteration of the LOT algorithm was run on the 897 eligible patients. All complex cases who, according to the algorithm output, received more than one LOT and/or maintenance therapy were automatically selected for chart review. The chart review identified 246 patients as complex cases, and then 254 patients who had only a single LOT and never received maintenance therapy (simple cases) were randomly chosen to complete the sample of 500 patients. The 500 cases were then split into training and test cohorts with the same ratios of simple:complex cases (Figure 2).

No significant differences in patient characteristics were found between the training and test cohorts (Table 1).

Figure 2. Selection of 500 patients whose deidentified charts were included in the study. LOT: line of therapy; NSCLC: non–small cell lung cancer.

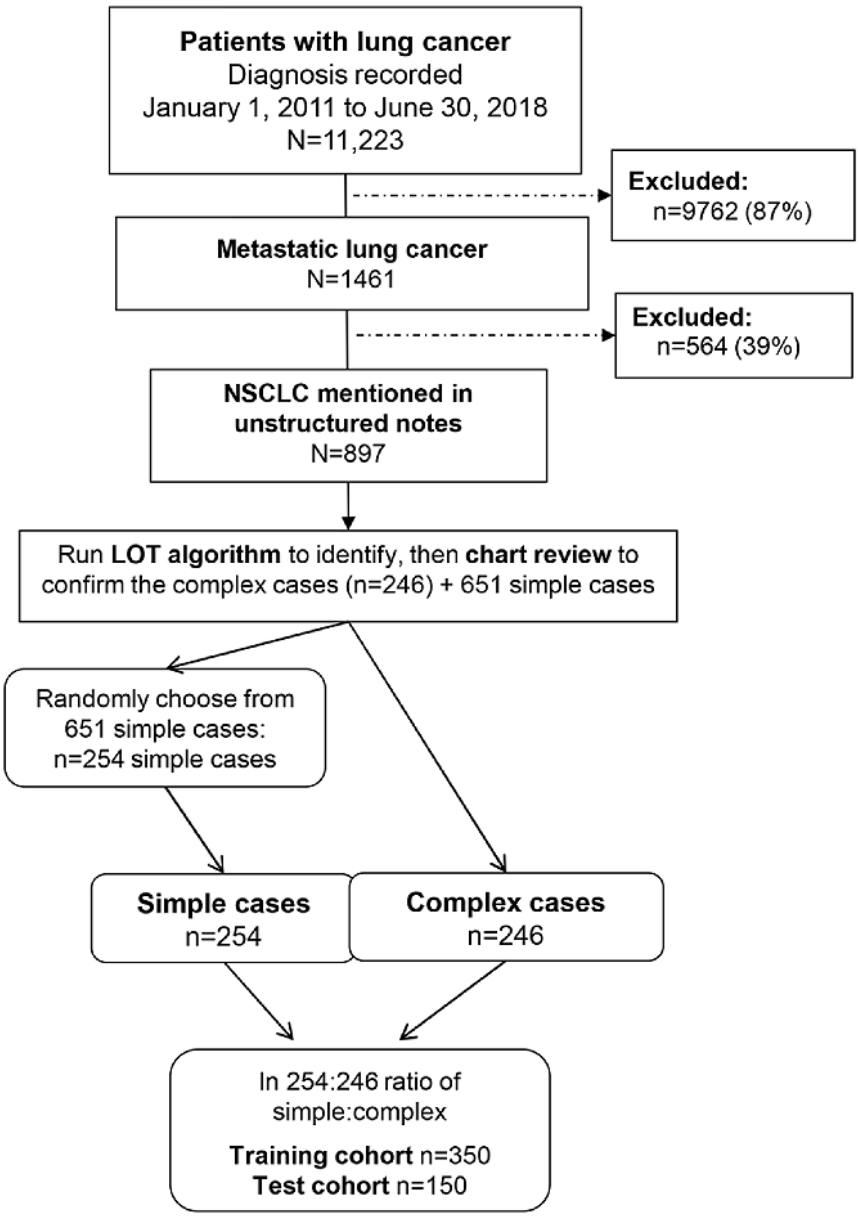


Table 1. Patient demographic characteristics.

Demographics	All patients (N=500)	Training cohort (n=350)	Test cohort (n=150)	*P* value
Female, n (%)	220 (44.0)	153 (43.7)	67 (44.7)	.85[a]
Age (years)				.16[b]
Mean (SD)	64.3 (10.7)	64.8 (10.7)	63.3 (10.5)	
Range	25-91	34-91	25-90	
Race,[c] n (%)				.95[a]
White	442 (89.1)	308 (88.8)	134 (89.9)	
Black	50 (10.1)	36 (10.4)	14 (9.4)	
Asian	4 (0.8)	3 (0.9)	1 (0.7)	
Hispanic, Latino, or other ethnicity	2 (0.4)	1 (0.3)	1 (0.7)	.49[a]

[a]Fisher exact test comparing training and test cohorts.

[b]Linear model analysis of variance comparing training and test cohorts.

[c]No information on race was available for 4 patients.

Blinded Manual Chart Review Findings

The distributions of LOT counts were similar between the training and test cohorts, and simple and complex cases each represented approximately half of the cases (Table 2).

A total of 55.1% (193/350) patients in the training cohort received one LOT. An additional 29.4% (103/350) had two LOTs, and 10.3% (36/350) had three LOTs. Most patients had three or fewer LOTs during their treatment history, and 14.6% (51/350) patients received maintenance therapy (Table 2).

Table 2. Blinded manual chart review findings for lines of therapy.

Group	All patients (N=500)	Training cohort (n=350)	Test cohort (n=150)	*P* value
Case classification, n (%)[a]				N/A[b]
Simple cases	254 (50.8)	178 (50.9)	76 (50.7)	
Complex cases	246 (49.2)	172 (49.1)	74 (49.3)	
LOT,[c] n (%)				.09[d]
1	280 (56.0)	193 (55.1)	87 (58.0)	
2	144 (28.8)	103 (29.4)	41 (27.3)	
3	51 (10.2)	36 (10.3)	15 (10.0)	
4	20 (4.0)	17 (4.9)	3 (2.0)	
5	3 (0.6)	0 (0.0)	3 (2.0)	
6	2 (0.4)	1 (0.3)	1 (0.7)	
Maintenance therapy, n (%)	74 (14.8)	51 (14.6)	23 (15.3)	.89[d]

[a]Blinded manual chart review was used to identify simple cases as patients who received one line of therapy (LOT) and no maintenance therapy, and complex cases as patients who received more than one LOT and/or maintenance therapy.

[b]N/A: not applicable.

[c]LOT: line of therapy.

[d]Fisher exact test comparing training and test cohorts.

Training Cohort: NSCLC LOT Algorithm Refinement

Overview

The ICCs on the number of LOTs between the LOT algorithm and chart review in the training cohort were 0.81 overall and 0.71 in the complex cases. The prerefinement agreement between the LOT algorithm output and chart review was 91% for the simple cases overall and 61% for the complex cases in the training cohort (Table 3).

Table 3. Intraclass correlation coefficients (ICCs) on number of lines of therapy (LOTs) and percentage agreement of non–small cell lung cancer LOT algorithm output with manual chart review.[a]

Group	Training cohort (n=350)		Test cohort (n=150)
	Prerefinement	Postrefinement	
Overall,[b] ICC[c] (95% CI)	0.81 (0.77-0.84)	0.87 (0.84-0.89)	0.90 (0.86-0.92)
Complex cases, ICC (95% CI)	0.71 (0.63-0.78)	0.75 (0.68-0.81)	0.82 (0.73-0.88)
Number of LOTs,[d] percentage agreement (95% CI)			
1 (train n=193, test n=87)	88.6 (84.1-93.1)	93.3 (89.7-96.8)	93.1 (87.8-98.4)
2 (train n=103, test n=41)	68.0 (58.9-77.0)	72.8 (64.2-81.4)	56.1 (40.9-71.3)
3 (train n=36, test n=15)	58.3 (42.2-74.4)	58.3 (42.2-74.4)	53.3 (28.1-78.6)
4 (train n=17, test n=3)	23.5 (3.4-43.7)	23.5 (3.4-43.7)	33.3 (0.0-86.7)
5 (train n=0, test n=3)	—[e]	—	—
6 (train n=1, test n=1)	—	—	—
Simple cases, overall	91.0 (86.8-95.2)	95.5 (92.5-98.5)	100
Complex cases, overall	60.5 (53.2-67.8)	64.0 (56.8-71.1)	50.0 (38.6-61.4)

[a]Simple or complex designation and mutually exclusive groups based on the total number of lines of therapy according to the chart review.

[b]The overall intraclass coefficients included data for simple cases, whereas simple cases were not evaluated separately because of low variability.

[c]ICC: intraclass coefficient.

[d]LOT: line of therapy.

[e]Patient numbers for five and six LOTs were too few for analysis.

For the simple cases, we found that the majority of discordances reflected the LOT not being advanced by chart review but being advanced in the algorithm output because of the 120-day gap-in-therapy rule (Table 4). A minor source of discordance was a difference in the LOT name, specifically when an initial drug was administered but then quickly dropped. Another minor source of discordance involved the 28-day line regimen defining window, that is, when a drug included in an LOT by chart review was not captured in the algorithm output because it had just missed the 28-day window.

Table 4. Reasons for discordance between the non–small cell lung cancer line-of-therapy algorithm and blinded chart review: numbers of cases.[a]

Reason for discordance	Training cohort (n=350)				Test cohort (n=150)	
	Prerefinement		Postrefinement		n (%)	N
	n (%)	N	n (%)	N		
Simple cases, total discordance	16 (9.0)	178	8 (4.5)	178	0 (0)	76
Gap-in-therapy window length	9 (56)	16	3 (38)	8		
28-day line regimen window	3 (19)	16	3 (38)	8		
Line name disagreement	3 (19)	16	1 (13)	8		
Other	1 (6)	6	1 (13)	8		
Complex cases, total discordance	68 (39.5)	172	62 (36)	172	37 (50.0)	76
Dropped drugs	22 (32)	68	24 (39)	62	17 (46)	37
Maintenance therapy classification	14 (21)	68	13 (21)	62	8 (22)	37
28-day line regimen window	12 (18)	68	12 (19)	62	6 (16)	37
Gap-in-therapy window length	9 (13)	68	4 (6)	62	2 (5)	37
Other	11 (16)	68	9 (15)	62	4 (11)	37

[a]Percentages may not add up to 100 because of rounding.

For the complex cases, the most common source of discordance resulted from dropped drugs, specifically cases when chart review advanced the LOT after a drug in combination therapy was dropped, but the algorithm did not (Table 4). For example, combination therapy with pembrolizumab-carboplatin followed by dropping carboplatin would trigger a new LOT of pembrolizumab monotherapy by chart review but not in the algorithm output.

The second most common source of discordance occurred because of differences in the identification of maintenance therapy, such as the determination of maintenance therapy after L1 regimens by chart review but not in algorithm output. Chart review often labeled maintenance therapy beyond L1 and/or with drugs outside the National Comprehensive Cancer Network (NCCN) guidelines, whereas the algorithm identified maintenance therapy in the L1 setting and using a drug list defined by NCCN guidelines [20,21]. Another reason for maintenance therapy discordance was whether the introduction of a new drug constituted a switch maintenance therapy or a new LOT. We did not attempt to resolve these discordances because of the subjective nature of any decision defining which method produced the true definition of maintenance therapy.

Discordances related to the 120-day gap-in-therapy window and the 28-day regimen window were also relatively common among the complex cases (Table 4).

Refinements Made to the NSCLC LOT Algorithm and Results of Refinement

After reviewing the discordances between chart review findings and LOT algorithm output for the training cohort, we used descriptive statistics and plots to determine how to adjust the discordant parameters and improve concordance when possible. For example, we identified the need to increase the gap-in-therapy window from 120 to 180 days by plotting the gap between successive prescriptions, excluding several protein kinase inhibitors as exceptions to the rule for gap-in-therapy line advancement (these *-tinib* drugs target tumor mutations such as *EGFR* and *ALK* genomic aberrations). In addition, we added gemcitabine as a continuation maintenance therapy, implemented the ability to advance the line if a drug in combination therapy were dropped, and implemented the ability to ignore drugs that were dropped during the 28-day line regimen-defining window (Table 5).

Table 5. Line-of-therapy algorithm parameters for metastatic non–small cell lung cancer: prerefinement and postrefinement.

Basic modules	Parameters	
	Prerefinement	Postrefinement
L1[a] first drug	On or after index date[b]	On or after index date[b]
Line regimen window	≤28 days after first drug	≤28 days after first drug
New drug line advancement	First instance	First instance
Exceptions (allowed substitutions)	Cisplatin ↔ carboplatin or paclitaxel ↔ albumin-bound paclitaxel substitution	Cisplatin ↔ carboplatin or paclitaxel ↔ albumin-bound paclitaxel substitution
Gap in therapy window	>120 days	>180 days
Exceptions (allowed gaps)	None	Erlotinib, afatinib, brigatinib, crizotinib, ceritinib, alectinib, gefitinib, osimertinib
Additional modules		
Maintenance therapy drugs		
Continuation maintenance	Bevacizumab, pemetrexed, atezolizumab	Bevacizumab, pemetrexed, atezolizumab, gemcitabine
Switch maintenance	Pemetrexed, docetaxel	Pemetrexed, docetaxel
Combination dropped drugs to advance LOT[c]	N/A[d]	Optional flag (not implemented)[e]
Drug switch during initial regimen window	N/A	Optional flag (not implemented)

[a]L1: first line of therapy.

[b]Index date defined as date of recorded metastatic non–small cell lung cancer diagnosis.

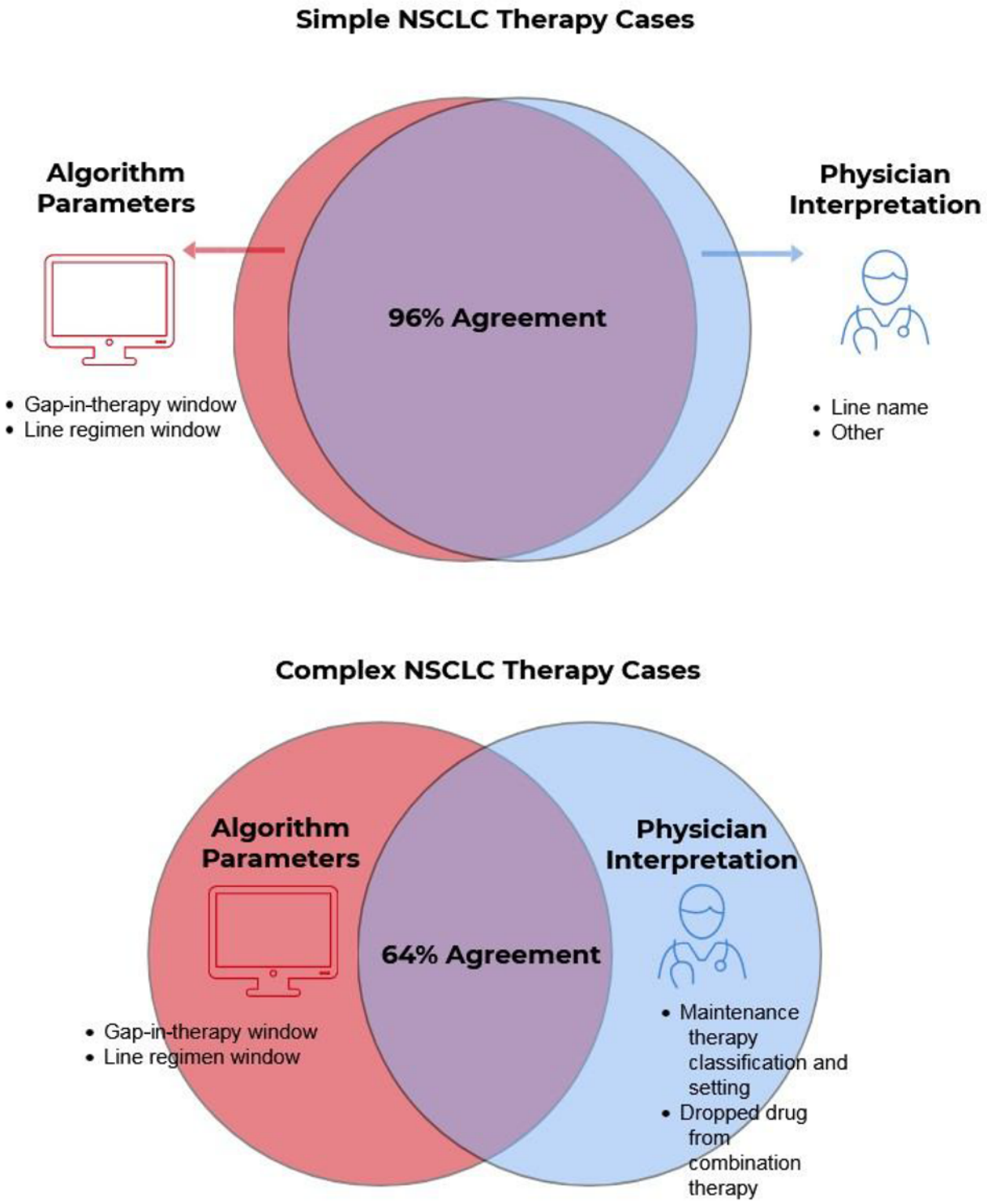
[c]LOT: line of therapy.

[d]N/A: not applicable.

[e]Option included in LOT to handle these cases but not used in this study.

Postrefinement agreement increased from 91% to 96% for the simple cases overall and from 61% to 64% for the complex cases, although improvements were limited to receipt of one or two LOTs (Table 3; Figure 3). In addition, postrefinement ICCs increased from 0.81 to 0.87 overall and 0.71 to 0.75 in the complex cases after refinement.

Figure 3. Results of applying the line-of-therapy algorithm to the training cohort postrefinement. NSCLC: non–small cell lung cancer.



After the LOT algorithm was refined, the total number of discordant results was halved for the simple cases in the training cohort, with the greatest decrease in discordance resulting from the increase from 120 to 180 days in the gap-in-therapy window (Table 4). For the complex cases in the training cohort, discordant numbers decreased only from 68 prerefinement to 62 postrefinement, with most of the decrease resulting from the change in the gap-in-therapy window. However, the number of cases with dropped drugs increased, indicating that fixing one issue can create other issues.

Test Cohort: Results for the NSCLC LOT Algorithm

The LOT algorithm was then run for the test cohort. For the simple cases, the agreement between the chart review results and algorithm output was 100% (Table 3), with no discordance (Table 4). For the complex cases, agreement was 50% overall and there were 37 instances of discordance, most commonly

because of differences resulting from dropped drugs, a pattern similar to that seen for the training cohort. The ICCs were 0.90 overall and 0.82 in the complex cases.

For patients who received one LOT, agreement was high and improved slightly with algorithm refinement (89% prerefinement, 93% postrefinement, 93% test cohort; Table 3). We observed a large decrease in agreement for patients who received more than one LOT.

Discussion

Principal Findings

We found an overall good alignment between our automated method of LOT classification and blinded manual chart review. As expected, the likelihood of precise matching between LOT algorithm output and chart review regarding LOT and maintenance therapy identification decreased with an increasing number of unique SACT regimens. This finding is consistent with the simple compounding of errors, that is, the chance of at least one error being found in multiple LOTs is greater than finding an error in a single LOT. On a per-LOT basis, the error would presumably remain fairly constant.

For the purposes of our comparisons, we used manual chart review as the gold standard. We improved the concordance between the LOT algorithm and chart review by increasing the gap-in-therapy window from 120 to 180 days. Concordance was also improved by adding drug class exceptions for protein kinase inhibitors to the gap-in-therapy rule and by adding gemcitabine as a continuation maintenance candidate. Our study notably contributes to the literature insofar as it identifies common sources of discordance between an LOT algorithm and clinician documentation, providing for the possibility of targeted algorithm refinement.

Addressing Areas of Discordance

Our study is one of the first to validate and refine an open-source LOT algorithm using manual chart review [22]. We note that although there were other potential opportunities to improve the percentage agreement between the algorithm and chart review, we could not identify clear recommendations in the medical literature or among experts to support modifications. Three areas of discordance with the potential to improve agreement included the following: (1) the decision whether to advance the LOT if a drug in a combination regimen is dropped, (2) whether a maintenance therapy could be classified as such beyond the first-line setting, and (3) whether a new drug administration during an LOT constitutes a line advancement or switch maintenance therapy. Although we treated manual chart review as the gold standard, clinical notes do not always document SACT administration in strict accordance with the definitions of LOT and maintenance therapy. Moreover, clinicians may disagree on classifications, leaving room for interpretation.

In the case of (1) whether to advance the line if a drug in a combination regimen is dropped, particular drugs may be dropped because of adverse events. Whether the remaining drugs should be considered the original or a new SACT regimen (LOT) is a subjective matter and may not be explicitly recorded by the prescribing physician. In the case of issues (2) and (3), NCCN guidelines specify that maintenance therapy is prescribed in the first-line setting and that a prescribed set of drugs is eligible for switch maintenance therapy for NSCLC [20,21]. However, these guidelines are not always followed, and maintenance drugs can be prescribed in an atypical manner. Moreover, maintenance therapy is often not recorded as such in clinical notes.

These small apparent inconsistencies may reflect a lack of precise definitions in LOT classification rules, or perhaps more likely, that physicians are instead appropriately focused on dynamically selecting optimal SACT regimens for their patients rather than precisely categorizing LOT and maintenance therapy. In addition, as shown in Tables 3 and 4 by complex cases not improving as the LOT numbers increased, refinements to the algorithm can create other inconsistencies when looking at the entire record. For example, after postrefinement, the algorithm agreed with chart review after exempting *-tinib* drugs from the 180-day rule but added an additional disagreement that resulted from changing the discontinuation gap from 120 to 180 days. Therefore, inconsistencies in physician LOT and maintenance classification, as well as algorithmic edge cases, make it unlikely that an automated LOT algorithm will achieve 100% alignment with independent chart review. In recognition of these unavoidable inconsistencies in LOT classification, we leave many parts of the algorithm to be highly configurable based on the specific use cases of researchers. We anticipate that other groups will make other choices with respect to configuration settings, but our study helps clarify the relative importance of these configuration settings.

Our algorithm is adaptable for use with other cancers and other cancer stages because of its modular design [6]. For example, drug lists, treatment sequences, and temporal parameters, such as the length of the gap between treatments, can be adjusted as appropriate for other tumor types and stages.

Study Limitations

This study has some limitations. First, we did not consider the length of oral drug administration. Oral drugs are often prescribed with a preset supply, and because the last dose administration is not typically recorded, extrapolation would be needed to determine the length of administration. In this study, our agreement metrics accounted for only the LOT number and regimen, and not the LOT duration; therefore, we considered only the first dose of oral drugs. We note also that we purposely oversampled for complex cases; therefore, the metrics reflect a distribution of patients that was not representative of the overall distribution. For example, only 28% of our selected study population versus 60% of eligible patients in the database received just one LOT without maintenance therapy, our definition of a simple case. Therefore, it is possible that single LOT metrics are under-represented. Finally, it could have been helpful to have more than one physician conducting the manual chart reviews, with an additional independent physician to resolve any discrepancies or disagreements.

Further research is needed on other data sets to determine if the results and conclusions are generalizable. In addition,

considerations such as detecting drug cycles and accounting for drug-specific nuances may increase the robustness of the algorithm. Further research on the appropriate metrics and benchmarks may be needed to address issues such as error compounding.

Conclusions

This study validates an EHR- and claims-based algorithm using medical chart review. We have refined the algorithm, highlighted areas of discordance, and noted the error compounding on further lines, allowing a deeper understanding of how the LOT algorithm may be used. We envision contributions to different disease indications and areas. In addition, common data set benchmarks, metrics, and increased accessibility will contribute substantially toward the development and adoption of this tool. Finally, a database of specific business rules concerning individual drugs and other nuanced behaviors will increase the robustness of the algorithm.

Acknowledgments

The authors thank Elizabeth V Hillyer, DVM, for editorial assistance. This work was supported by Merck Sharp & Dohme Corporation, a subsidiary of Merck & Company, Inc, Kenilworth, NJ, United States.

Conflicts of Interest

WM and WO are employees of Merck Sharp & Dohme Corp, a subsidiary of Merck & Co, Inc, Kenilworth, NJ, USA, and stockholders of Merck & Co, Inc, Kenilworth, NJ, USA. KMM, KAL, and PRD have no conflicts of interest to declare. ARR and AG are employees of Regenstrief Institute, which was paid by Merck Sharp & Dohme Corp to conduct this study.

Multimedia Appendix 1
Identification of patients with metastatic non–small cell lung cancer.
[DOCX File , 13 KB - medinform_v9i10e29017_app1.docx]

References

1. Hirsch FR, Scagliotti GV, Mulshine JL, Kwon R, Curran WJ, Wu Y, et al. Lung cancer: current therapies and new targeted treatments. Lancet 2017 Jan 21;389(10066):299-311. [doi: 10.1016/S0140-6736(16)30958-8] [Medline: 27574741]
2. Brody H. Lung cancer. Nature 2020 Nov;587(7834):S7. [doi: 10.1038/d41586-020-03152-0] [Medline: 33208969]
3. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Ann Glob Health 2019 Jan 22;85(1):8 [FREE Full text] [doi: 10.5334/aogh.2419] [Medline: 30741509]
4. Minguet J, Smith KH, Bramlage P. Targeted therapies for treatment of non-small cell lung cancer--Recent advances and future perspectives. Int J Cancer 2016 Jun 01;138(11):2549-2561 [FREE Full text] [doi: 10.1002/ijc.29915] [Medline: 26537995]
5. Reck M, Rabe KF. Precision diagnosis and treatment for advanced non-small-cell lung cancer. N Engl J Med 2017 Aug 31;377(9):849-861. [doi: 10.1056/NEJMra1703413] [Medline: 28854088]
6. Meng W, Ou W, Chandwani S, Chen X, Black W, Cai Z. Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer. J Biomed Inform 2019 Dec;100:103335 [FREE Full text] [doi: 10.1016/j.jbi.2019.103335] [Medline: 31689549]
7. Rajkumar SV, Harousseau J, Durie B, Anderson KC, Dimopoulos M, Kyle R, International Myeloma Workshop Consensus Panel 1. Consensus recommendations for the uniform reporting of clinical trials: report of the International Myeloma Workshop Consensus Panel 1. Blood 2011 May 05;117(18):4691-4695 [FREE Full text] [doi: 10.1182/blood-2010-10-299487] [Medline: 21292775]
8. Rossi A. New options for combination therapy for advanced non-squamous NSCLC. Expert Rev Respir Med 2019 Nov;13(11):1095-1107. [doi: 10.1080/17476348.2019.1667233] [Medline: 31512526]
9. Berger ML, Curtis MD, Smith G, Harnett J, Abernethy AP. Opportunities and challenges in leveraging electronic health record data in oncology. Future Oncol 2016 May;12(10):1261-1274. [doi: 10.2217/fon-2015-0043] [Medline: 27096309]
10. Lee W, Bridewell W, Das AK. Alignment and clustering of breast cancer patients by longitudinal treatment history. AMIA Annu Symp Proc 2011;2011:760-767 [FREE Full text] [Medline: 22195133]
11. Rocque GB, Kandhare PG, Williams CP, Nakhmani A, Azuero A, Burkard ME, et al. Visualization of sequential treatments in metastatic breast cancer. JCO Clin Cancer Inform 2019 Mar;3:1-8 [FREE Full text] [doi: 10.1200/CCI.18.00095] [Medline: 30840488]
12. Abernethy AP, Arunachalam A, Burke T, McKay C, Cao X, Sorg R, et al. Real-world first-line treatment and overall survival in non-small cell lung cancer without known EGFR mutations or ALK rearrangements in US community oncology setting. PLoS One 2017 Jun 23;12(6):e0178420 [FREE Full text] [doi: 10.1371/journal.pone.0178420] [Medline: 28644837]
13. Bittoni MA, Arunachalam A, Li H, Camacho R, He J, Zhong Y, et al. Real-world treatment patterns, overall survival, and occurrence and costs of adverse events associated with first-line therapies for Medicare patients 65 years and older with

advanced non-small-cell lung cancer: a retrospective study. Clin Lung Cancer 2018 Sep;19(5):629-645 [FREE Full text] [doi: 10.1016/j.cllc.2018.04.017] [Medline: 29885945]
14. Arunachalam A, Li H, Bittoni MA, Camacho R, Cao X, Zhong Y, et al. Real-world treatment patterns, overall survival, and occurrence and costs of adverse events associated with second-line therapies for Medicare patients with advanced non-small-cell lung cancer. Clin Lung Cancer 2018 Sep;19(5):783-799 [FREE Full text] [doi: 10.1016/j.cllc.2018.05.016] [Medline: 29983370]
15. Ramsey SD, Martins RG, Blough DK, Tock LS, Lubeck D, Reyes CM. Second-line and third-line chemotherapy for lung cancer: use and cost. Am J Manag Care 2008 May;14(5):297-306 [FREE Full text] [Medline: 18471034]
16. Grannis SJ, Stevens KC, Merriwether R. Leveraging health information exchange to support public health situational awareness: the indiana experience. Online J Public Health Inform 2010;2(2):3213 [FREE Full text] [doi: 10.5210/ojphi.v2i2.3213] [Medline: 23569586]
17. Dixon BE, Grannis SJ, McAndrews C, Broyles AA, Mikels-Carrasco W, Wiensch A, et al. Leveraging data visualization and a statewide health information exchange to support COVID-19 surveillance and response: application of public health informatics. J Am Med Inform Assoc 2021 Jul 14;28(7):1363-1373 [FREE Full text] [doi: 10.1093/jamia/ocab004] [Medline: 33480419]
18. Ruppert LP, He J, Martin J, Eckert G, Ouyang F, Church A, et al. Linkage of Indiana State Cancer Registry and Indiana network for patient care data. J Registry Manag 2016;43(4):174-178. [Medline: 29595920]
19. Line of therapy algorithm (open source). GitHub. URL: https://github.com/Merck/Line-of-Therapy-Algorithm [accessed 2021-08-10]
20. Ettinger DS, Wood DE, Aggarwal C, Aisner DL, Akerley W, Bauman JR, OCN [Corporate Author], et al. NCCN guidelines insights: non-small cell lung cancer, version 1.2020. J Natl Compr Canc Netw 2019 Dec;17(12):1464-1472. [doi: 10.6004/jnccn.2019.0059] [Medline: 31805526]
21. NCCN clinical practice guidelines in oncology: non-small cell lung cancer. National Comprehensive Cancer Network. URL: http://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf [accessed 2021-08-10]
22. Nordstrom BL, Simeone JC, Malley KG, Fraeman KH, Klippel Z, Durst M, et al. Validation of claims algorithms for progression to metastatic cancer in patients with breast, non-small cell lung, and colorectal cancer. Front Oncol 2016 Feb 1;6:18 [FREE Full text] [doi: 10.3389/fonc.2016.00018] [Medline: 26870695]

Abbreviations

ALK: anaplastic lymphoma kinase
ANOVA: analysis of variance
EGFR: epidermal growth factor receptor
EHR: electronic health record
L1, L2, L3, L4: first-, second-, third-, and fourth-line therapy
LOT: line of therapy
NCCN: National Comprehensive Cancer Network
NSCLC: non–small cell lung cancer
SACT: systemic anticancer therapy

Edited by C Lovis; submitted 22.03.21; peer-reviewed by T Burke, C Zeng, P Ray; comments to author 17.05.21; revised version received 22.06.21; accepted 02.07.21; published 12.10.21.

Please cite as:
Meng W, Mosesso KM, Lane KA, Roberts AR, Griffith A, Ou W, Dexter PR
An Automated Line-of-Therapy Algorithm for Adults With Metastatic Non–Small Cell Lung Cancer: Validation Study Using Blinded Manual Chart Review
JMIR Med Inform 2021;9(10):e29017
URL: https://medinform.jmir.org/2021/10/e29017
doi:10.2196/29017
PMID:34636730

Original Paper

Privacy-Preserving Anonymity for Periodical Releases of Spontaneous Adverse Drug Event Reporting Data: Algorithm Development and Validation

Jie-Teng Wang[1*], MSc; Wen-Yang Lin[1*], PhD, Prof Dr

Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan

[*]all authors contributed equally

Corresponding Author:
Wen-Yang Lin, PhD, Prof Dr
Department of Computer Science and Information Engineering
National University of Kaohsiung
700 Kaohsiung Univ. Rd, Nanzih District
Kaohsiung, 811
Taiwan
Phone: 886 7 5919517
Fax: 886 7 5919514
Email: wylin@nuk.edu.tw

Abstract

Background: Spontaneous reporting systems (SRSs) have been increasingly established to collect adverse drug events for fostering adverse drug reaction (ADR) detection and analysis research. SRS data contain personal information, and so their publication requires data anonymization to prevent the disclosure of individuals' privacy. We have previously proposed a privacy model called MS(k, θ*)-bounding and the associated MS-Anonymization algorithm to fulfill the anonymization of SRS data. In the real world, the SRS data usually are released periodically (eg, FDA Adverse Event Reporting System [FAERS]) to accommodate newly collected adverse drug events. Different anonymized releases of SRS data available to the attacker may thwart our single-release-focus method, that is, MS(k, θ*)-bounding.

Objective: We investigate the privacy threat caused by periodical releases of SRS data and propose anonymization methods to prevent the disclosure of personal privacy information while maintaining the utility of published data.

Methods: We identify potential attacks on periodical releases of SRS data, namely, BFL-attacks, mainly caused by follow-up cases. We present a new privacy model called PPMS(k, θ*)-bounding, and propose the associated PPMS-Anonymization algorithm and 2 improvements: PPMS+-Anonymization and PPMS++-Anonymization. Empirical evaluations were performed using 32 selected FAERS quarter data sets from 2004Q1 to 2011Q4. The performance of the proposed versions of PPMS-Anonymization was inspected against MS-Anonymization from some aspects, including data distortion, measured by normalized information loss; privacy risk of anonymized data, measured by dangerous identity ratio and dangerous sensitivity ratio; and data utility, measured by the bias of signal counting and strength (proportional reporting ratio).

Results: The best version of PPMS-Anonymization, PPMS++-Anonymization, achieves nearly the same quality as MS-Anonymization in both privacy protection and data utility. Overall, PPMS++-Anonymization ensures zero privacy risk on record and attribute linkage, and exhibits 51%-78% and 59%-82% improvements on information loss over PPMS+-Anonymization and PPMS-Anonymization, respectively, and significantly reduces the bias of ADR signal.

Conclusions: The proposed PPMS(k, θ*)-bounding model and PPMS-Anonymization algorithm are effective in anonymizing SRS data sets in the periodical data publishing scenario, preventing the series of releases from disclosing personal sensitive information caused by BFL-attacks while maintaining the data utility for ADR signal detection.

(JMIR Med Inform 2021;9(10):e28752) doi:10.2196/28752

KEYWORDS

adverse drug reaction; data anonymization; incremental data publishing; privacy preserving data publishing; spontaneous reporting system; drug; data set; anonymous; privacy; security; algorithm; development; validation; data

Introduction

Motivation

Adverse drug reactions (ADRs) are undesirable side effects of taking drugs. Before hitting the market, a new drug has to undergo a series of clinical trials. Unfortunately, it is hard to find all ADRs in the premarketing stage due to fewer volunteers. Thus, an increasing number of countries have built spontaneous reporting systems (SRSs) to collect adverse drug events (ADEs) to monitor the safety of marketed drugs, such as the FDA Adverse Event Reporting System (FAERS) of the US Food and Drug Administration (FDA) [1], the UK Yellow Card scheme [2], and the MedEffect Canada [3]. Some countries even publish their SRS data sets, for example, US FDA and MedEffect Canada, to the public to facilitate ADR research.

SRS data are a kind of microdata containing personal health information, such as diseases of the patients. Microdata, usually represented in the form of tables of tuples [4], are composed of explicit identifier (*ID*) that can uniquely identify each individual (eg, SSN, name, phone number); quasi-identifier (*QID*) that can be linked with external data to reidentify some of the individuals (eg, sex, age, and ZIP code); sensitive attribute (*SA*) that contains sensitive information, such as disease or salary; and non-SA that falls into none of the above 3 categories. Publishing these data sets would lead to privacy threats. A real case did occur in Canada. A broadcaster successfully reidentified a 26-year-old girl by linking MedEffect Canada and the publicly available obituaries [5]. This case motivated the research by El Emam et al [5], whose findings showed that the MedEffect Canada data exhibit a high risk of identity disclosure.

Generally, simple removal of the identification attributes, such as name, SSN, or phone, has been shown to fail to protect individual privacy [6]. The adversary can still link published data to external data (eg, voter list, through quasi-identification attributes, such as gender, job, age, ZIP code). This calls for the research topic, namely, privacy-preserving data publishing (PPDP), which aims to anonymize raw data before publication. In [7], we pointed out that none of traditional anonymization methods (eg, *k*-anonymity [6], *l*-diversity [8]) is favorable for SRS data sets due to characteristics such as multiple individual records, multivalued SAs, and rare events. Later, we proposed a privacy model called MS(k, θ*)-bounding [9] to anonymize SRS data to prevent the disclosure of individual privacy. New events arrive in SRSs continuously in the real world, so countries such as the USA and Canada release SRS data sets periodically, for example, every quarter, to handle this kind of dynamically growing data sets (ie, periodical data publishing). Unfortunately, MS(k, θ*)-anonymity is designed for a single static publishing scenario, and is awkward to handle a series of published data sets.

Usually, each ADE record in SRS data contains a CaseID to trace the follow-ups of that event; all records with the same CaseID, located within the same or different periods, refer to the same event. Although someone may regard follow-ups as duplicates of the original case, the situation is somewhat different. Follow-up cases contain complement or correction of the original case. Still, duplicate reports refer to the same case submitted by different reporters, so were misrecorded with different CaseIDs. Follow-ups are easily detected via CaseID, but identifying actual duplicates is challenging, which should be considered a data preprocessing issue. There has been some research studies on detecting actual duplicates in SRS data [10-12]. Most SRS systems such as FAERS, however, provide no deduplication mechanism. We thus ignore this issue. Unfortunately, CaseID provides a useful linkage for the adversary across a series of anonymized data sets to exclude records not belonging to the target, raising the risk of breaching the target's privacy. For illustration, let us consider 3 consecutive quarters of published SRS data sets in Table 1, each of which satisfies 3-anonymity.

Table 1. Three consecutive quarters of published spontaneous reporting system data sets, each satisfying 3-anonymity.

Quarter and CaseID	Sex	Age	Disease
1			
1	Male	[35-40]	Flu
2	Male	[35-40]	Flu
3	Male	[35-40]	Fever
4	Female	[30-35]	HIV
5	Female	[30-35]	Flu
6	Female	[30-35]	Diabetes
2			
1	ANY	[30-40]	Flu
4	ANY	[30-40]	HIV
7	ANY	[30-40]	Diabetes
8	Male	[30-35]	Fever
9	Male	[30-35]	Flu
10	Male	[30-35]	Diabetes
11	Male	[30-35]	HIV
12	Male	[30-35]	Flu
3			
13	Female	[30-35]	Flu
14	Female	[30-35]	Diabetes
15	Female	[30-35]	Fever
16	Female	[30-35]	Flu
17	Female	[30-35]	Fever
7	Male	[30-35]	Diabetes
8	Male	[30-35]	Fever
18	Male	[30-35]	HIV

Possible Scenarios

Scenario I

Suppose that the adversary learns that his/her neighbor Alice, whose *QID* value is {Female, 32}, suffered from some ADR in Q2. First, the adversary links to Table 1 (quarter 2) through the *QID* of Alice, learning that the record of Alice is in the first *QID* group (CaseIDs 1, 4, and 7). The adversary can then link to the previously published SRS data through the candidate CaseID set {1, 4, 7} and find the record with CaseID=1 and Sex=Male in Table 1 (quarter 1). Because Alice is female, the adversary can exclude CaseID 1 from the candidate CaseID set {1, 4, 7}, changing Table 1 (quarter 2) to 2-anonymous and lifting the confidence of the attacker to identify Alice.

Scenario II

Following the previous example, the adversary has known the candidate CaseID set of Alice {4, 7}. The adversary can now use this set to link to subsequently published SRS data and observe a record whose CaseID is 7 in Table 1 (quarter 3). Because the owner of that record is male, the adversary can exclude CaseID 7 from the candidate CaseID set, concluding that the CaseID of Alice in Table 1 (quarter 2) is 4.

Scenario III

Suppose that the adversary learns John's *QID* value is {Male, 33} and the first time that John had an ADR is in Q3. This means that the CaseID of John's event is a "new CaseID" in Q3 and shall not appear in any previously released data. First, the adversary links to Quarter 3 and learns that the record of John is within the second *QID* group (CaseIDs 7, 8, 18). The adversary can then connect to the 2 previously published SRS data sets through the candidate CaseID set of John {7, 8, 18}, observing 2 matching records whose CaseID are 7 and 8 in Quarter 2. The CaseID of John is neither 7 nor 8, so the adversary concludes that the CaseID of John is 18, ruining the privacy protection embedded by 3-anonymity.

Background Knowledge and Related Work

Privacy Models for Microdata Publishing

Research on PPDP [4] aims to protect released microdata from 2 types of privacy attacks: *record disclosure* and *attribute disclosure*.

Record disclosure, also known as *table linkage attack*, refers to the situation in which the individual identity of a specific tuple that has been deidentified in the published data is reidentified. Although it is hard to prevent table linkage attacks, it is possible to reduce the possibility of identifying victims in a published data. Achievement is the invention of *k*-anonymity [6], which is the most influential privacy model that generalizes the values of *QID* to ensure that each record in published data contains at least *k*–1 other records with the same *QID* value.

Attribute disclosure, also known as *attribute linkage attack*, refers to the situation in which attackers can infer an individual's sensitive information, even though they fail to perceive the exact record of the victim. Unfortunately, *k*-anonymity is not able to prevent attribute disclosure. Another renowned privacy model called *l*-diversity [8] was thus proposed. The main idea of *l*-diversity is to thwart the adversary's belief on the probability of the sensitive value by ensuring that each *QID* group contains at least *l* "well-represented" sensitive values, that is, the probability of inferring the sensitive value of the victim will be at most 1/*l*.

Privacy Models for Incremental Data Publishing

Most real-world data are not static but dynamically changing, which means that data cannot be published simultaneously but have to be published incrementally [4]. Previously proposed privacy models such as *k*-anonymity and *l*-diversity only focus on single static data publishing, awkward to prevent privacy disclosure in incremental data publishing. Contemporary privacy models for incremental data publishing can be classified into *continuous* or *dynamic* data publishing [4].

Continuous Data Publishing

This refers to the scenario in which all data collected so far have to be published even if some of the data have been released before. More precisely, suppose that the data holder had previously collected a set of records D_1 time stamped t_1 and published the anonymized version R_1 of D_1. After collecting a new set of records D_2 time stamped t_2, the data holder will publish R_2 as an anonymized version of all records collected so far, (ie, $D_1 \cup D_2$). In general, the published release R_i ($i \geq 1$) shall be an anonymized version of $D_1 \cup D_2 \cup ... D_i$.

Byun et al [13] first identified the privacy threat under continuous data publishing. They demonstrated possible inference channels by comparing different *l*-diverse releases to explore the sensitive values of victims. They later enhanced their approach by considering both *k*-anonymity and *l*-diverse called (*k*, *c*)-anonymous and exploring more types of adversarial attacks named *cross-version inference*s [14].

Pei et al [15] illustrated that in the continuous data publishing scenario, the adversary can infer some privacy information from multiple releases that have been sanitized by *k*-anonymity. They also proposed an effective method called "monotonic incremental anonymization," which would progressively and consistently reduce the generalization granularity as the updates arrive to maintain *k*-anonymity.

Fung et al [16] proposed a method to quantify the exact number of records that can be "cracked" by comparing the series of published *k*-anonymous data. The adversary can exclude the cracked records from published data, making the published data no longer satisfy *k*-anonymous. They also presented a privacy model, called *BCF*-anonymity, to measure the anonymous number in published data after excluding the cracked records, and proposed an algorithm to anonymize published data achieving *BCF*-anonymity.

Dynamic Data Publishing

This refers to the scenario in which the data holder can insert records into or delete records, or perform both actions, from raw data sets. Suppose that the data holder had collected an initial set of records D_1 in time t_1 and published its anonymized version R_1. During the period [t_1, t_2), the data holder kept collecting new records and inserted them into D_1. Further, the data holder might delete and update some records from D_1, finally obtaining the updated version D_2 of D_1 in t_2. Then, the published release R_2 in t_2 is an anonymized version of D_2. In general, a published release R_i in time t_i shall be an anonymized version of D_i.

Xiao and Tao [17] identified a kind of privacy disclosure called *critical absence*. The adversary can infer victims' sensitive information by comparing the series of published *l*-diverse data in dynamic data publishing scenarios (only considered insertion and deletion). They proposed a privacy model, called *m*-invariance, to ensure the certain "invariance" of the "signature" of *QID* groups, and an effective method called counterfeited generalization to anonymize published data achieving *m*-invariance.

Bu et al [18] noticed that some sensitive values would be permanent, such as criminal record and some incurable diseases, such as HIV. They showed that *m*-invariance is unable to prevent privacy disclosure when permanent sensitive values are present. Therefore, they proposed an anonymization approach, called *HD*-composition [18], to limit the probability of linkage between individuals and sensitive values not over a given threshold.

On observing *m*-invariance only considers data evolution caused by insertion and deletion, Li and Zhou [19] further presented a counterfeit generalization model named *m*-distinct to support full data evolution (ie, insertion, update, and deletion). Moreover, they observed that attribute updates are seldom arbitrary, with some correlations often existing between the old and the new values. Based on this observation, they assumed that all updates on sensitive values are nonarbitrary. Therefore, *m*-distinct applies the concept of the candidate update set, which is a set of specific sensitive values that can be updated.

Following the work in [19], Anjum et al [20] further assumed that the updates in fully dynamic data publishing are arbitrary, meaning the old values of attributes may not correlate with the new values. They presented a new kind of attack named τ-attack by exploiting the "event list" of an individual. They also proposed a method called τ-safety, an extension of *m*-invariance, to solve the privacy disclosure caused by τ-attack.

He et al [21] presented a new type of attack named *value equivalence attack*, which can exploit the partitioned structure of published data, such as *m*-invariant releases, to obtain

sensitive information of individuals. Once the adversary knows the actual sensitive value of an individual, he/she can disclose the sensitive information of the remaining individuals within the same equivalence class. They proposed a graph-based anonymization algorithm, which leverages a min-cut algorithm to prevent the old "value association attack" and the new "equivalence attack."

Specifically, Bewong et al [22] focused on transactional data. They proposed a new privacy model called *serially preserving*, which requires the posterior probability of any sensitive term to its corresponding population rate bounded by a given threshold. A novel anonymization method (Sanony, which counts on adding counterfeits) was presented to guarantee a new published transactional data set satisfying the required privacy model.

There is another scenario of nonstatic data publishing called *sequential data publishing*. Different vertical projections of the same table on different subsets of attributes are published consecutively in this scenario. Anonymization models and methods for this scenario were first studied in [23] and then further investigated in [24] and [25].

In summary, no contemporary work notices the scenario of periodical data publishing, and no work has been conducted for SRS data anonymization, considering the privacy threat caused by follow-up cases. In this paper, we investigate the privacy threat caused by periodical releases of SRS data and propose anonymization methods to prevent the disclosure of personal privacy information while maintaining the utility of published data.

Methods

Publishing Scenario and Privacy Attacks

We first introduce the periodical data publishing scenario and present 3 kinds of privacy attacks for periodically published SRS data sets satisfying MS(k, θ*)-bounding. We propose a new privacy model, PPMS(k, θ*)-bounding, to protect published SRS data sets from those attacks in the periodical data publishing scenario. We also propose a corresponding anonymization algorithm, namely PPMS-anonymization, that incorporates 2 innovative strategies, *NC*-bounding and *QID*-covering, to prevent the released data sets from privacy attacks caused by follow-up key (ie, CaseID). Two extensions of PPMS-anonymization, PPMS+-anonymization and PPMS++-anonymization, are presented as well, which employ more efficient techniques, including neglecting subsequent coverings and grouping with new cases.

BFL-Attacks

Typical SRS data, such as FAERS, are usually published periodically and contain follow-up cases, which can be expressed as a new data publishing model named periodical data publishing. Suppose that the data holder previously had collected an initial set of records D_1 in period $[t_0, t_1)$ and published R_1 as an anonymized version of D_1. After collecting a new set of records D_2 during period $[t_1, t_2)$ the attacker wants to anonymize and publish D_2 at time t_2. D_2 may or may not contain some follow-up cases in D_1. Let R_2 denote the anonymized version of D_2. In general, the release R_i published at t_i is an anonymized version of D_i ($i \geq 1$). Note that for an original case x, the life span of its follow-up cases in subsequent releases is not continuous. That is, a follow-up observed in D_i may disappear in D_{i+1} but show up again in some later release D_{i+j}, for $j>1$. This makes the periodical publishing scenario distinct from existing scenarios in the literature. First, unlike the situation in dynamic data publishing, D_i is a new set of collections, rather than updated from D_{i-1}. Besides, the existence of follow-up cases is different from the assumption for continuous data publishing (ie, all cases in D_i should be kept in all subsequent releases D_j, for $j>i$). A comparison of the proposed periodical data publishing with dynamic data publishing and sequential data publishing is summarized in Multimedia Appendix 1 (also see Textbox 1).

Textbox 1. Definition 1: QID-cover.

Consider the *QID* values, q_1 and q_2, of 2 cases. We say q_1 covers q_2, denoted by $q_1 \sqsupseteq q_2$, if for every attribute a in *QID*, $a(q_1)$ is equal to or more generalized than $a(q_2)$, where $a(q)$ denotes the value of q in attribute a.

Backward-Attack (B-Attack)

Backward-Attack (*B*-attack) focuses on excluding records from the specific release by exploiting some previous ones (Textbox 2). Scenario I is an example, which occurs when the *QID* value of the old case differs from the background learned by the attacker. As the *QID* values would have been generalized in all published releases, the only way by which *B*-attack can succeed is when the *QID* value of old CaseID fails to cover that of the current CaseID. More precisely, for every target v, if in any previous release there exists an old CaseID i_{old} corresponding to the candidate CaseID set of v such that the *QID* value of i_{old} does not cover the *QID* value of v, then i_{old} would be excluded from the candidate CaseID set of v.

Textbox 2. Definition 2: Backward-attack.

Consider a target v to be inferred by the attacker and an anonymized release R_i. Let q^v and *CI* denote the *QID* value and the candidate CaseID set of v in R_i, respectively, and U be the set of records in all previous releases $\{R_1, R_2, ..., R_{i-1}\}$ whose CaseID is in *CI*. The *B*-attack will occur if there exists a record r in U such that the *QID* value of r, q^r, does not cover q^v. The set of these excludable records is denoted by *B*.

Forward-Attack (F-Attack)

Analogous to *B*-attack, Forward-Attack (*F*-attack) occurs when the *QID* value of the following CaseID differs from the background learned by the attacker (Textbox 3). That is, the *QID* value of a following CaseID in some subsequent releases fails to cover that of the current CaseID. An example is shown in Scenario II. More precisely, for every target v, if in any subsequent release there exists a following CaseID i_{new} corresponding to the candidate CaseID set of v such that the *QID* value of i_{new} does not cover the *QID* value of v, then i_{new} would be excluded from the candidate CaseID set of v.

Textbox 3. Definition 3: Forward-attack.

Consider a target v and an anonymized release R_i. Let q^v and *CI* denote the *QID* value and the candidate CaseID set of v in R_i, respectively, and U be the set of records in all subsequent releases $\{R_{i+1}, R_{i+2}, ..., R_c\}$ whose CaseID is in *CI*. The *F*-attack will occur if there exists a record r in U such that the *QID* value of r, q^r, does not cover q^v. The set of these excludable records is denoted by *F*.

Latest-Attack (L-Attack)

This attack is illustrated in Scenario III. In this example, the attacker knows that the event for the target (John) first appears in Quarter 3. It follows that John's case (CaseID) is definitely absent in all previously published releases. In general, for every target v whose CaseID is first present in some release known by the attacker, *Latest Attack* (*L*-attack) would occur if the candidate CaseID set of v contains some old CaseIDs appearing in previous releases (Textbox 4).

Textbox 4. Definition 4: Latest-attack.

Consider a target v. Suppose the attacker learns that the CaseID of v first appears in an anonymized release R_i. Let *CI* be the candidate CaseID set of v in R_i. The *L*-attack will occur if there exists any case in *CI* whose CaseID appears in some previous releases. The set of these excludable records is denoted by *L*.

Privacy Model PPMS(k, θ*)-bounding

To prevent *BFL*-attacks, we propose a new privacy model called periodical-publishing multisensitive (k, θ*)-bounding, abbreviated as PPMS(k, θ*)-bounding (Textboxes 5 and 6).

Textbox 5. Definition 5: Confidence.

Let s be a sensitive value in *SA* and an anonymized release R_i. Given a target v with *QID* value q^v, we define the probability that v has sensitive value s as $conf(v \rightarrow s)$, which is equal to $\sigma_s(g)/|g|$, where g denotes the *QID* group in R_i in which v resides and $\sigma_s(g)$ is the number of cases in g that contains s.

Textbox 6. Definition 6: PPMS(k, θ*)-bounding.

Let $S=\{s_1, s_2, ..., s_m\}$ be the set of all possible sensitive values in *SA* and $\theta^*=(\theta_1, \theta_2, ..., \theta_m)$ be the probability thresholds specified by the data holder, where $0 \le \theta_j \le 1$, for $1 \le j \le m$. We say a series of anonymized releases $R_1, R_2, ..., R_n$ satisfies PPMS(k, θ*)-bounding if each R_i, $1 \le i \le n$, satisfies the following:

1. For every individual v, the size of the candidate CaseID set *CI* of v in R_i excluding *B*, *F*, and *L* is no less than k, that is, $|CI - (B \cup F \cup L)| \ge k$, and

2. The confidence to infer v having any sensitive value $s_j \in S$ is no larger than θ_j, that is, $conf(v \rightarrow s_j) \le \theta_j$.

The privacy requirement of Definition 6(1) is to prevent record disclosure while Definition 6(2) is to prevent attribute disclosure. Our model adopts nonuniform thresholds for different sensitive values because different values express different degrees of sensitivity in the real world. For example, the disclosure of a patient with fever is far less sensitive than that of an individual with HIV.

Anonymization Algorithm

Overview

Our algorithm can be summarized as a greedy and clustering approach to divide records into *QID* groups. Viewing each *QID* group as a cluster, we adopted a clustering-based method [26] to build *QID* groups, each of which starts from a randomly chosen record and grows gradually by adding a solo record exhibiting the best characteristic among all candidates. This process repeats until the *QID* group satisfies the "k" requirement. Finally, we generalize the *QID* values of all records within the same cluster to the same value.

We adopted 2 metrics, information loss [26] (Textbox 7) and privacy risk (PR) [9] (Textbox 8), to choose the best isolated record. For each evolving *QID* group, the former favors the new record contributing minimal impact to the data utility while the latter quantifies the ratio of sensitive values within the *QID* group to meet the privacy requirement in Definition 6(2).

Textbox 7. Definition 7: Information loss.

Suppose the *QID* attributes can be separated to 2 different sets, numerical attributes $\{N_1, N_2, ..., N_m\}$ and categorical attributes $\{C_1, C_2, ..., C_n\}$, and each C_i is associated with a taxonomy tree T_i. Let g denote a *QID* group (or cluster). The *information loss* (*IL*) [26] of g is defined as follows:

where max(N_i) and min(N_i) denote the maximum and minimum values of attribute N_i in the whole data set, and max(N_i, g) and min(N_i, g) denote the maximum and minimum values of attribute N_i in g. Notation $|g|$ is the number of records in g, $h(C_j)$ the height of the taxonomy tree T_j, and $h(C_j, g)$ is the height of the generalized value of C_j in g in taxonomy tree T_j.

To find a new record r to be included in g, we choose the one causing the least increase of information loss, which is measured by

$$\Delta IL(g, r) = IL(g \cup \{r\}) - IL(g) \quad \textbf{(2)}$$

Then, the most feasible choice r_{bst} is

$$r_{bst} = \text{argmin}_r \, \Delta IL(g, r) \quad \textbf{(3)}$$

In addition, the inclusion of record r containing sensitive value s that appears in g would cause the ratio of s in g to be over θ_s. As we will derive in Lemma 2, we have to keep the occurrence of s in g, denoted by $\sigma_s(g)$, under a maximum threshold $\eta_s(g)$ to prevent the confidence of inferring sensitive value s in g from being larger than θ_s. We thus adopt the PR_s introduced in [9].

When $\eta_s(g \cup \{r\}) \geq \sigma_s(g \cup \{r\})$, a greater σ_s leads to a larger PR_s. Therefore, Equation 4 favors the new record r whose sensitive values are relatively rare in g. Because a record may contain more than 1 sensitive value, the PR caused by adding r into g can be defined as the summation of PR_s over all sensitive values.

Textbox 8. Definition 8: Privacy risk.

Let g denote a *QID* group (or cluster) during the execution of our anonymization algorithm. The PR [9] of adding a new record r into g is

where $s \in S_r$ and S_r is the set of sensitive values contained in record r.

The value of summation of PR_s may be zero, that is, all sensitive values in r are new to group g. An increment is thus added into PR(g, r) in Equation 5 to avoid zero PR. The smaller the PR caused by adding r into g, the more likely r will be chosen. If the inclusion of r makes the number of records containing s in g more than the maximally allowed number, PR becomes infinite, so r will not be chosen. Finally, we refine ΔIL into $\Delta IL'$ as follows

$$\Delta IL'(g, r) = \Delta IL(g, r) \times \text{PR}(g, r) \quad \textbf{(6)}$$

and the most feasible choice r_{bst} is

$$r_{bst} = \text{argmin}_r \, \Delta IL'(g, r) \quad \textbf{(7)}$$

Strategies Against BFL-Attacks

The *NC*-bounding strategy aims to maintain at least "*k*" new CaseID records in each group after excluding all old CaseID records. This is because all old CaseID records may become excludable by exploiting the previous releases, such as *B*-attack and *L*-attack. *QID*-covering is to generalize the *QID* value of records to prevent them from being excluded by *B*-attack and *F*-attack. *NC*-bounding allows the adversary to discover and exclude records not belonging to the target, but enforces the privacy requirement met by the remaining records. *QID*-covering, by contrast, perplexes the adversary to find out excludable records.

Strategy for L-Attack

Overview

Recall that *L*-attack occurs as the adversary knows the exact published release to which the first ADE of the target *v* belongs. Specifically, let this release be R_i. All old CaseIDs in target *v*'s *CI* set in R_i refer to other targets, which are potentially excluded by the attacker and so should be discounted from forming a valid *QID* group, that is, the size of the *QID* group should be at least *k*. For this reason, we use strategy *NC*-bounding.

Example 1

Consider the example in Scenario III. The target *QID* group <Male, [30-35]> in Table 1 (quarter 3) contains 2 old CaseIDs (ie, 7 and 8). We need to add 2 other records with new CaseIDs to make Table 1 (quarter 3) invulnerable to *L*-attack. In this case, all records in the *QID* group <Female, [30-35]> are new cases and the size of <Female, [30-35]> is larger than $k + 2$. We can choose any 2 of them (eg, 16 and 17) into <Male, [30-35]> and generalize the *QID* values accordingly. In general, to defend against *L*-attack, the number of new CaseID records in every *QID* group needs to be no less than *k*.

Strategy for B-Attack

Overview

Suppose the target v is in R_i. *B*-attack means the adversary can link to R_1, R_2, ..., R_{i-1} through the candidate CaseID set of v to exclude those CaseIDs definitely not belonging to target v. Note that all of the excludable CaseIDs in *B*-attack are old CaseIDs; thus, the situation is the same as *L*-attack in which all of the old CaseID records have a probability to be excluded. Therefore, the *NC*-bounding strategy used to defend *L*-attack can also be used to secure against *B*-attack. That is, the number of new CaseID records in every *QID* group needs to be larger than or equal to k in PPMS(k, θ*)-bounding. In this sense, *L*-attack is similar to *B*-attack, because both of them exploit the previous releases to find excludable CaseIDs. The main difference is that the former needs to know whether the CaseID is old or not, while the latter needs to compare the *QID* values to infer whether the CaseID belongs to the target.

Example 2

Consider the example in Scenario I. Similar to the previous example for *L*-attack, we have to include 2 records with new CaseIDs, say 8 and 9, into the *QID* group containing old CaseIDs 1 and 4 in Table 1 (quarter 2), that is, <ANY, [30-40]>, and perform generalization accordingly. Table 2 (quarter 2) shows the resulting anonymized table.

Table 2. The anonymized releases against *BFL*-attack for the example in Table 1.

Quarter and CaseID	Sex	Age	Disease
2			
13	Female	[30-35]	Flu
14	Female	[30-35]	Diabetes
15	Female	[30-35]	Fever
16	ANY	[30-40]	Flu
17	ANY	[30-40]	Fever
7	ANY	[30-40]	Diabetes
8	ANY	[30-40]	Fever
18	ANY	[30-40]	HIV
3			
1	ANY	[30-40]	Flu
4	ANY	[30-40]	HIV
7	ANY	[30-40]	Diabetes
8	ANY	[30-40]	Fever
9	ANY	[30-40]	Flu
10	Male	[30-35]	Diabetes
11	Male	[30-35]	HIV
12	Male	[30-35]	Flu

Strategy for F-Attack

Overview

Suppose the target is in R_i. *F*-attack means that the adversary can link to $\{R_{i+1}, R_{i+2}, ..., R_n\}$ through the candidate CaseID set of target and exclude the CaseIDs that are definitely not referring to the target. Unlike *BL*-attacks, *F*-attack exploits the subsequent releases. The *NC*-bounding strategy works for *BL*-attacks because we can find out which CaseIDs are excludable in the latest raw data set by using previous releases. Unfortunately, because R_{i+1}, R_{i+2}, ..., R_n is not published yet, there is no way to foresee which CaseIDs will be excluded in R_i by employing *F*-attack, causing the *NC*-bounding strategy to be infeasible to defend *F*-attack. By contrast, we know that the adversary can exploit R_i to perform *F*-attack to exclude records in R_1, R_2, ..., R_{i-1}. Therefore, the focus is to protect R_1, R_2, ..., R_{i-1} from *F*-attack through utilizing R_i. In other words, we have to consider how to anonymize D_i to R_i, making R_i non-exploitable for performing *F*-attack on R_1, R_2, ..., R_{i-1}. By applying the same strategy to all subsequent releases after R_i, that is, R_{i+1}, R_{i+2}, ..., R_n, we protect R_i from *F*-attack.

Let OC_i be the set of old CaseIDs present in at least one of the previous releases R_1, R_2, ..., R_{i-1}. Consider a record r whose CaseID is in OC_i. Let $O=\{r_1, r_2, ..., r_p\}$ refer to, as in previous releases R_1, R_2, ..., R_{i-1}, the set of records that has the same CaseID as that of r. To prevent *F*-attack, we have to ensure that

$$\forall a \in QID, a(r) \not\equiv a(r_i), \text{ for } 1 \geq i \geq p.$$

That is, the *QID* value of r should cover that of all r's previous cases.

Example 3

Consider the example in Scenario II. To prevent the table published in Quarter 2 from *F*-attack, we have to generalize the 2 records, 7 and 8, in Quarter 3 to cover their corresponding predecessors in Table 1 (quarter 2). This causes the *QID* value of case 7 to become "ANY, [30-40]" and that of case 8 remains unchanged. Because 7, 8, and 18 are in the same *QID* group, we have to generalize their *QID* values into the same value, that is, "ANY, [30-40]". Finally, if *L*-attack is considered as well, as demonstrated in Example 1, we have to include cases 16 and 17 and finally obtain the result in Table 2 (quarter 2).

Lemma 1 (Covering Transitivity)

Consider any 3 records, r_1, r_2, and r_3, with the same CaseID in 3 anonymous releases R_i, R_j, and R_k, $i<j<k$. If $q^{r1} \sqsubseteq q^{r2}$ and $q^{r2} \sqsubseteq q^{r3}$, then $q^{r1} \sqsubseteq q^{r3}$.

Lemma 1 suggests an efficient approach for realizing *QID* covering against *F*-attack. When we are anonymizing D_i to R_i, rather than checking all of the old CaseID records in the previous releases, $\{R_1, R_2, ..., R_{i-1}\}$, we only have to search for, starting from R_{i-1} to R_1, the latest release containing old CaseID records. Once we find that release, we can stop checking the remaining ones.

We next summarize how we can integrate these 2 strategies to meet the privacy requirement in Definition 6(a).

Theorem 1

A release R_i anonymized by following strategies of *NC*-bounding and *QID* covering satisfies the requirement of Definition 6(a). For proof, please see Multimedia Appendix 2.

Strategy Against Attribute Disclosure

Overview

The privacy disclosure caused by *BFL*-attacks not only includes record disclosure but also attribute disclosure. This is illustrated with the following example.

Example 4

Consider the 3 consecutive quarters of the 3-anonymous release in Table 1. Recall that in Scenario I the adversary can link to Table 1 (quarter 3) through the *QID* value of Alice {Female, 32} and perceive the *CI* of Alice is {1, 4, 7}, inferring the probability of Alice having any of {Flu, HIV, Diabetes} is 1/3. After employing *B*-attack via Quarter 1, *CI* is reduced to {4, 7}, so the adversary's confidence that Alice has HIV or diabetes increases to 1/2. He/she can further exclude CaseID 7 from *CI* by performing *F*-attack via Quarter 3 and be 100% sure that Alice has HIV.

Now let us consider how to prevent the attribute disclosure caused by *BFL*-attacks. The basic idea is to control the ratio of sensitive values in each *QID* group to be no greater than the specified threshold. Consider our proposed strategies against *BFL*-attacks stated in the previous section. Let $S_g=\{s_1, s_2, ..., s_p\}$ denote the set of sensitive values in g and $(\theta_1, \theta_2, ..., \theta_p)$ the corresponding threshold specified for S_g. We can derive the following occurrence bound for each sensitive value within a *QID* group g to meet the required threshold.

Lemma 2

For any sensitive value $s \in S_g$, the maximal number of cases in g that contains s without breaking the associated threshold θ_s, denoted by $\eta_s(g)$, is

where $|NC(g)|$ is the number of new CaseIDs in g. For proof, please see Multimedia Appendix 3.

Algorithm PPMS-Anonymization

Multimedia Appendix 4 presents our algorithm PPMS-Anonymization, which is composed of 3 stages. The first stage aims at finding out old CaseID records and generalizing their *QID* values in advance to achieve *QID*-covering against *F*-attack. Because there may exist multiple individual records [9] in ADE data sets, we follow the *combined record* (or *super record*) concept in [9] to deal with this issue. All records with the same CaseID are combined into a super record before starting to form *QID* groups. Without this process, the records with identical CaseIDs may be divided into different *QID* groups, leading to more substantial deviation in the data quality and perplexing the process of identifying duplicate records while detecting ADR signals.

To find out old CaseID records in D_i and generalize their *QID* values in advance, we check previous releases R_{pre} from R_{i-1} to R_{i-x} (if i=1, R_{pre}=null). Because CaseID is used to trace an event's follow-ups, there is typically a life span of CaseID, denoted by x. The generalization of old CaseID records aims at achieving *QID*-covering against *F*-attack. Because of the transitive property of *QID* value shown in Lemma 1, once we discover an old CaseID record r' in any one of the previous releases, we stop checking the remaining earlier releases by using "break" (line 13 in Multimedia Appendix 4) to end the "while loop" (line 8 in Multimedia Appendix 4).

The second stage shown in Multimedia Appendix 5 is activated by calling the procedure *Grouping*, forming as many *QID* groups satisfying PPMS(k, θ*)-bounding as possible. Each group begins with a randomly chosen seed record, gradually growing by adding a record with the least $\Delta IL'$ (defined in Equation 7) until there are at least k new CaseID records to achieve the *NC*-bounding strategy. The *OldCaseNum* function returns the number of old CaseID records in a group. A new group then begins with the new record most distinguished from the one just added into the latest group. The above steps are repeated until the remaining records fail to form a group, for example, the number of new CaseID records is less than k or the ratio of all sensitive values within the remaining records is higher than the associated threshold (see line 10 in Multimedia Appendix 5).

The last stage is activated by calling the function *Generalization* (Multimedia Appendix 6), which processes the remaining ungrouped records by assigning each of them into the most feasible group that produces the minimal $\Delta IL'$ to sustain the

data utility and satisfy the privacy requirement. Next, the super records will be split back to the original records (the group they belong to remains unchanged). Finally, all records within the same group are generalized into the same *QID* value to satisfy PPMS(k, θ*)-bounding.

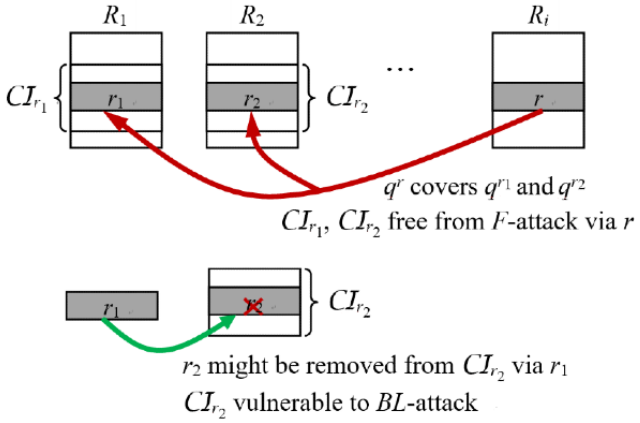
Algorithm PPMS$^+$-Anonymization

In this section, we propose an improvement of our PPMS-Anonymization algorithm: PPMS$^+$-Anonymization. The idea is to neglect the *QID* covering derived in Lemma 1.

Let r be a record in D_i whose CaseID is c, q^r the *QID* value of r, and r_1, r_2, ..., r_p be the older versions of r in the previous releases R_1, R_2, ..., R_{i-1}. To prevent *F*-attack, we have to make q^r cover $\{q^{r1}, q^{r2}, ..., q^{rp}\}$. Although we have exploited the transitivity property in Lemma 1 to avoid checking out all of the old CaseID records in releases R_1, R_2, ..., R_{i-1}, the *QID* value suffers from accumulated generalization. That is, the later the record r is published, the more information loss will be caused by generalization. Fortunately, we can limit the accumulated generalization by neglecting all subsequent *QID* coverings.

The fact is that some of the records protected by *QID*-covering against *F*-attack still can be eliminated by *BL*-attacks. Following the previous discussion, let r_1 be the earliest record with CaseID=c. Without loss of generality, assume r_1 resides in R_1. Then clearly, c is a new case in R_1, that is, $c \in NC(R_1)$, and will be an old case in all subsequent releases, that is, $c \in OC(R_j)$, $2 \le j \le i-1$. Remember that all old CaseIDs have the potential to be excluded by *BL*-attacks. So even if we make q^r cover $\{q^{r2}, q^{r3}, ..., q^{ri-1}\}$ to prevent $\{r_2, r_3, ..., r_{i-1}\}$ from being excluded by *F*-attack, they can still be excluded by *BL*-attack. This means that generalizing q^r to cover $\{q^{r2}, q^{r3}, ..., q^{ri-1}\}$ is useless. It suffices to generalize q^r to cover q^{r1}. Figure 1 illustrates this concept.

Figure 1. Idea illustration of neglecting subsequent coverings.



Multimedia Appendix 7 shows PPMS$^+$-Anonymization, the improved version of PPMS-Anonymization in Multimedia Appendix 4 (lines 5-18). For the given record r, the modified version seeks R_{i-x} to R_{i-1} to find the earliest release in which r occurs. Once we find out the earliest old CaseID record r', we stop checking the remaining releases.

Algorithm PPMS^{++}-Anonymization

Overview

In Multimedia Appendix 5, the procedure Grouping works by picking and adding the record with the least $\Delta IL'$ into the group, overlooking whether the record is a new or an old case in D'. We observed that this mixture of new and old cases to form a *QID* group would paralyze the discrimination of ΔIL in choosing good candidate records, that is, Equation 7, and cause severe information loss.

Suppose an old CaseID record r is picked as the seed to start a new *QID* group g in the procedure Grouping. As an old case, the *QID* value of r has already been generalized to cover its earliest clone record r' in some previous release, meaning that q^r is as coarser as the group in which r' resides. Therefore, if there exist some isolated records whose *QID* values are covered by q^r, then adding these records into g yields no increase in information loss (ie, ΔIL=0). Although this does not affect the information loss of group g, it does increase the information loss of the selected record. And in this situation, the Grouping procedure will randomly choose one from those isolated records, disregarding different degrees of information loss brought to these isolated records.

Example 5

Consider Table 3. We assume the age attribute has been discretized following the taxonomy tree in Multimedia Appendix 8. The first 3 records form a group starting with the old case record 1, while records 4, 5, and 6 are new cases. Adding any of the 3 isolated records into this group yields no change in the group information loss because all of their *QID* values are covered by record 1. This makes no distinction in choosing the isolated records, but record 6 is the best choice, which exhibits the least data distortion after *QID* generalization.

Table 3. An illustration of the problem of *QID* grouping starting with an old case.

QID group and isolated records	Sex	Age	Disease
A forming *QID* group			
CaseID 1	ANY	Nonadult	Flu
CaseID 2	ANY	Nonadult	Flu
CaseID 3	ANY	Nonadult	Fever
Isolated records			
CaseID 4	Female	Newborn	Fever
CaseID 5	Male	Preschool	Flu
CaseID 6	Female	Adolescent	Diabetes

To solve this problem, we avoid mixing new CaseID and old CaseID records in forming *QID* groups. Instead, we separate old CaseID records from *D* before starting the procedure Grouping, forming possible *QID* groups composed of only new CaseID records. The set of old CaseID records and the remaining new CaseID records are later dealt with by the function Generalization. Multimedia Appendix 9 describes the modification of Multimedia Appendix 4 to realize PPMS^{++}-Anonymization, an improvement of PPMS^{+}-Anonymization by grouping new cases first.

Results

Overview

We designed a series of experiments to examine the effectiveness of our new method in anonymizing a series of periodically released SRS data sets. The proposed PPMS-Anonymization algorithm and its extensions, PPMS^{+}-Anonymization and PPMS^{++}-Anonymization, were compared with method MS-Anonymization. In this section, we describe the details of each experiment, including the experimental results and our observations.

Experimental Setup

The data used in our experiment consist of 32 quarterly collections from FAERS, including 2004Q1 to 2011Q4. We used attributes {*Weight*, *Age*, *Gender*} as *QID*, where *Weight* is numerical while the other 2 are categorical, with drug indication (*INDI_PT*) and drug reaction (*PT*) as *SA*. To view *Age* as categorical, we adopted the age taxonomy defined in MeSH [27] (Multimedia Appendix 8). Moreover, we discarded records that have missing values in either *QID* or *SA* attributes.

We respectively performed MS-Anonymization [9] and 3 versions of PPMS-Anonymization, including the original version of PPMS-Anonymization (PPMS), the improved version by incorporating neglecting subsequent coverings (PPMS^{+}), and the advanced version by employing neglecting subsequent coverings and grouping with new cases (PPMS^{++}), to anonymize the selected FAERS data sets, and computed the information loss of 2 series of anonymized data sets. We then imitated the behavior of the adversary, employing *BFL*-attacks to find out all excludable CaseIDs in 2 series of anonymized data sets. After that, we removed all excludable records, and evaluated the risk of record and attribute disclosure of 2 series of anonymized data sets.

We examined 2 aspects of anonymized data sets: information loss and PR. The information loss of an anonymized data set is measured by *normalized information loss* (*NIL*), meaning the average *IL* (using Equation 1) for each attribute of each record.

where *R* is an anonymized data set, *g* is a *QID*-group, *GroupNum*(*R*) denotes the number of *QID* groups in *R*, and |*QID*| is the number of attributes in *QID*. This yields *NIL* ranging in [0-1]; the larger the *NIL* is, the more serious is the information loss.

We also used the 2 criteria in [9] to measure the privacy disclosure, *dangerous identity ratio* (*DIR*) and *dangerous sensitivity ratio* (*DSR*); the former measures the ratio of *QID* groups that violate the privacy requirement for protecting record identity, while the latter measures the ratio of *QID* groups that explore sensitive values.

$$DIR(R)=DIGNum(R)/GroupNum(R) \quad \textbf{(10)}$$

$$DSR(R)=DSGNum(R)/GroupNum(R) \quad \textbf{(11)}$$

If the number of records in a *QID* group is less than the threshold *k*, we say this group is a *dangerous identity group* (*DIG*). *DIGNum*(*R*) denotes the number of *DIG*s in the anonymized data set *R*. A *QID* group is a *dangerous sensitivity group* (*DSG*) if it contains at least one unsafe sensitive value whose frequency is higher than the associated threshold. *DSGNum*(*R*) denotes the number of *DSGs* in *R*.

To observe the influence of 2 anonymization methods on the strength of ADR signals, we chose from FDA MedWatch [28] all significant ADR rules involving patient demographics such as age or gender conditions and causing withdrawal or warning of the drug. A detailed description of these ADR rules is presented in Table 4. We used the proportional reporting ratio (PRR) [29] description (Multimedia Appendix 10) to measure the strength of ADR signals, which is used by the UK Yellow Card database and UK Medicines and Healthcare products Regulatory Agency (MHRA).

Table 4. Selected adverse drug reaction rules from Food and Drug Administration MedWatch.

Drug name and adverse reaction	Demographic condition	Marked year	Withdrawn or warning year
Avandia			
• Myocardial infarction • Death • Cerebrovascular accident	Age>18	1999	2010
Tysabri			
• Progressive multifocal leukoencephalopathy	Age>18	2004	2005
Zelnorm			
• Cerebrovascular accident	Sex=Female	2002	2007
Warfarin			
• Myocardial infarction	Age>60	1940	2014
Revatio			
• Death	Age>18	2008	2014

We considered 3 ways of setting θ*. First, we applied a uniform setting on θ*, that is, all confidence thresholds of symptoms were set to the same value (0.2 or 0.4). Then, we used a frequency-based method to determine the threshold of each symptom, which is based on the following idea: The more frequently the symptom occurs, the less sensitive it is. For this purpose, we calculated the average count of symptoms m and the corresponding SD. Then we set the confidence thresholds of symptoms whose occurrence is less than m – SD, between m – SD and m + SD, and higher than m + SD to 0.2, 0.6, and 1, respectively. Last, we adopted a level-wise confidence setting, which is similar to the frequency setting but conforming to well-recognized medical sensitive terms. All symptoms were classified into 3 levels: high sensitive (θ=0.2), low sensitive (θ=0.4), and nonsensitive (θ=1.0). For this purpose, we followed the setting in [9], choosing the group of symptoms related to AIDS: "Acquired immunodeficiency syndromes" in MedDRA (Medical Dictionary for Regulatory Activities) as high sensitive, 2 groups called "Coughing and associated symptoms" and "Allergies to foods, food additives, drugs and other chemicals" as nonsensitive, and those not belonging to the above groups as low sensitive.

Results on Anonymization Quality

This section will report the results on information loss and privacy disclosure of MS-Anonymization and our proposed 3 versions of PPMS-Anonymization under 3 different settings of θ*.

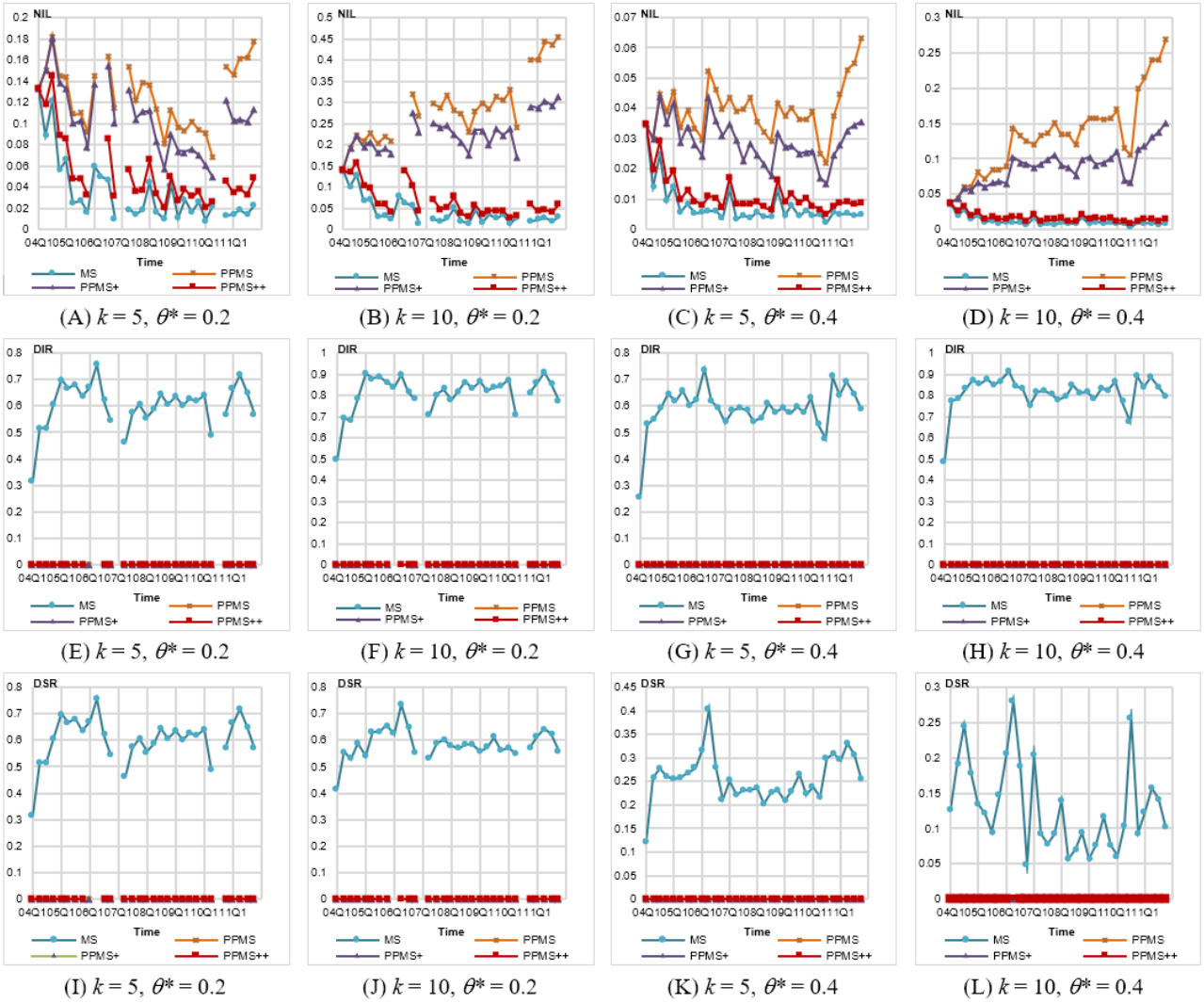
Uniform Confidence Setting

In this evaluation, we set a uniform threshold (θ*=0.2 and 0.4) to each symptom, that is, the sensitivity of each symptom is the same, and 2 settings of k (k=5, 10).

Information Loss

First, we evaluated the information loss. As per the results shown in Figure 2A-D, the general trend is when θ* is lower, the information loss is higher. It is because more records with different sensitive values have to be grouped together to form a valid *QID* group, so more generalization has to be performed. Among the 3 versions of PPMS-Anonymization, PPMS^{++} leads the rank, followed by PPMS^{+} and PPMS, with average improvements of 51% and 59% for PPMS^{++} over PPMS^{+} and PPMS, respectively, as θ*=0.2 and k=5, and reaching 78% and 82% for θ*=0.4 and k=10. We noticed that as θ*=0.2, some anonymized data sets fail to meet the privacy requirement, that is, 2006Q1, 2006Q2, 2007Q1, and 2010Q3. A further inspection revealed that these data sets contain some highly frequent symptoms. For example, there are 20,467 cases (without missing values) in 2007Q1, and 3877 (18.94%) of them contain "Diabetes Mellitus Non-Insulin-Dependent". All methods fail in this data set because the minimum bound of that symptom should be 21.00% (3877/18,462, where 18,462 is the number of new cases), so the privacy requirement of 20% cannot be satisfied. In the data set 2010Q3, there are 12,727/56,550 (22.51%) cases containing "Smoking Cessation Therapy," so no method can meet the privacy requirement. (In 2006Q1 and 2006Q2, the symptom "Myocardial Infarction" is frequent.) In general, the uniform threshold setting is not suitable, especially when some sensitive values are persistent.

Figure 2. Evaluation on information loss and privacy disclosure for Federal Drug Administration Adverse Event Reporting System (FAERS) data anonymized by different methods with uniform setting of θ*. DIR: dangerous identity ratio, DSG: dangerous sensitivity group, NIL: normalized information loss, PPMS: periodical-publishing multisensitive.



Record Disclosure

Next, we compared the record disclosure caused by each method. The results are shown in Figure 2E-H. MS-Anonymization exhibits serious record disclosure. The average *DIR*s for k=5 and 10 are 0.61 and 0.8, respectively, meaning over half of *QID* groups are *DIG*s. Besides, the *DIR* of MS-Anonymization increases as k is larger. This is because a larger k leads to less number of groups and so a higher ratio of groups containing old cases, increasing the risk of *QID* groups becoming dangerous. It is noteworthy that the *DIR*s of 3 versions of PPMS-Anonymization are all 0. The reason is that our method guarantees free of record disclosure and the *DIR* metric is not dependent on different settings of θ*.

Attribute Disclosure

Finally, we present the results on the *DSR* metric. The results are shown in Figure 2I and J. MS-Anonymization yields very high *DSR*s, 0.6 on average, for lower θ* values (θ=0.2). This is because a lower θ is more likely to cause the number of symptoms close to its maximal allowed number in the *QID* groups, especially for high-frequent symptoms. Thus, the action of excluding records is more likely to cause the violation of θ* and so leads to relatively higher *DSR*s, such as 2006Q1, 2006Q2, 2007Q1, and 2010Q3. For example, the maximal symptom frequencies in 2006Q4 and 2010Q1 are only 8.1% and 9.1%, respectively, relatively smaller than θ*=0.2 or 0.4, so the *DSR*s of these 2 releases are relatively lower than other releases. This again demonstrates that the uniform threshold setting is not feasible. The setting of k also influences the *DSR*s yielded by MS-Anonymization. A larger k not only causes higher maximal allowed numbers of symptoms in *QID* groups but also reduces the change in the ratio of symptoms when some records are excluded. Compared with MS-Anonymization, all 3 versions of PPMS-Anonymization yield zero *DSR* value in all data sets, except 2006Q1, 2006Q2, and 2007Q, showing our method can protect data from attribute disclosure caused by *BFL*-attacks.

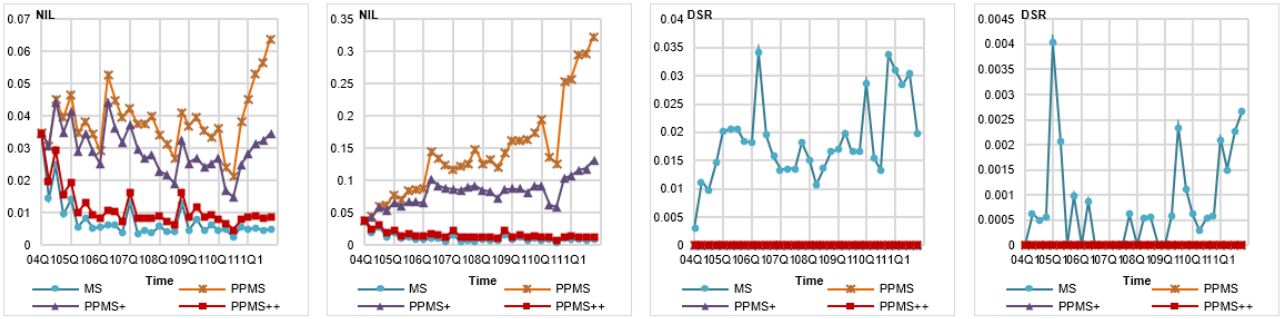
Frequency-Based Confidence Setting

Two different settings of k (5 or 10) are considered. The results on *DIR* are omitted because they are similar to those generated by uniform setting, that is, MS-Anonymization generates large *DIR*s while our PPMS-Anonymization yields zero *DIR*.

Information Loss

As shown in Figure 3A and B, the *NIL*s generated by each method are better than those under the uniform setting. It is not surprising because this more flexible setting easily allows the methods to choose the closer new record to be added during *QID* group construction. Similar to those observed for the uniform setting, PPMS^{++} significantly outperforms PPMS^{+} and PPMS, yielding *NIL*s less than 0.05 for k =5 and 0.15 for k =10.

Figure 3. Evaluation on information loss and privacy disclosure for Federal Drug Administration Adverse Event Reporting System (FAERS) data anonymized by different methods with frequency-based setting of θ*. DSR: dangerous sensitivity ratio, NIL: normalized information loss, PPMS: periodical-publishing multisensitive.



(A) $k = 5$ (B) $k = 10$ (C) $k = 5$ (D) $k = 10$

Attribute Disclosure

As shown in Figure 3C and D, all data sets anonymized by PPMS-Anonymization are free of attribute disclosure (ie, zero *DSR*). The *DSR*s of MS-Anonymization are very small compared with those in previous settings. It is because those *DSG*s in the previous experiments are caused by high frequent symptoms, whose thresholds, however, are set to 1 in this experiment. In FAERS data, there are more than 20,000 different symptoms. It is hard to determine a suitable threshold for each of them without background knowledge. Therefore, the frequency-based method is a convenient and reasonable way to deal with this issue. This also demonstrates the value of allowing nonuniform settings in our model.

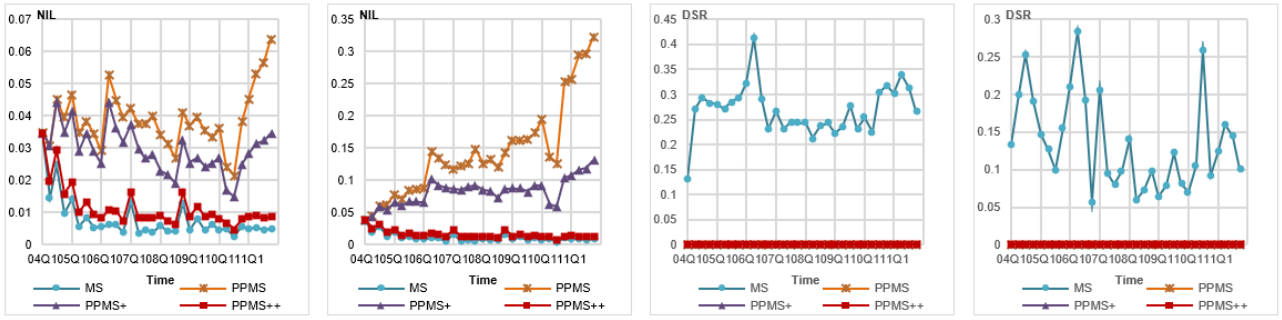
Level-Wise Confidence Setting

Again, 2 different k (5 and 10) settings are considered, and for the same reason, we omit the results on *DIR*.

Information Loss

Figure 4A and B shows that although PPMS and PPMS^{+} yield more information loss than that by MS-Anonymization, PPMS^{++} behaves comparably to MS-Anonymization. The *NIL*s are very similar to those under the frequency-based setting.

Figure 4. Evaluation on information loss and privacy disclosure for Federal Drug Administration Adverse Event Reporting System (FAERS) data anonymized by different methods with level-wise setting of θ*. DSR: dangerous sensitivity ratio, NIL: normalized information loss, PPMS: periodical-publishing multisensitive.



(A) $k = 5$ (B) $k = 10$ (C) $k = 5$ (D) $k = 10$

Attribute Disclosure

The results in Figure 4C and D show that all 3 versions of PPMS-Anonymization cause no attribute disclosure (with zero *DSR*s), but large *DSR*s are observed for MS-Anonymization. We can see that the *DSR*s of MS-Anonymization in some quarters are relatively higher, just similar to the results in Figure 2K and L and Figure 3C and D.

Influence on ADR Signals

Selected Signals

In this experiment, we inspected variation on the strength of observed ADR signals shown in Table 4 between before and after anonymization. Because some signals exhibit similar performance, we only show 3 representatives with different demographic conditions, that is, the signals related to Avandia, Zelnorm, and Warfarin, which are shown as follows:

R1: Avandia, Age>18 → Myocardial infarction

R2: Zelnorm, Sex=Female → Cerebrovascular accident

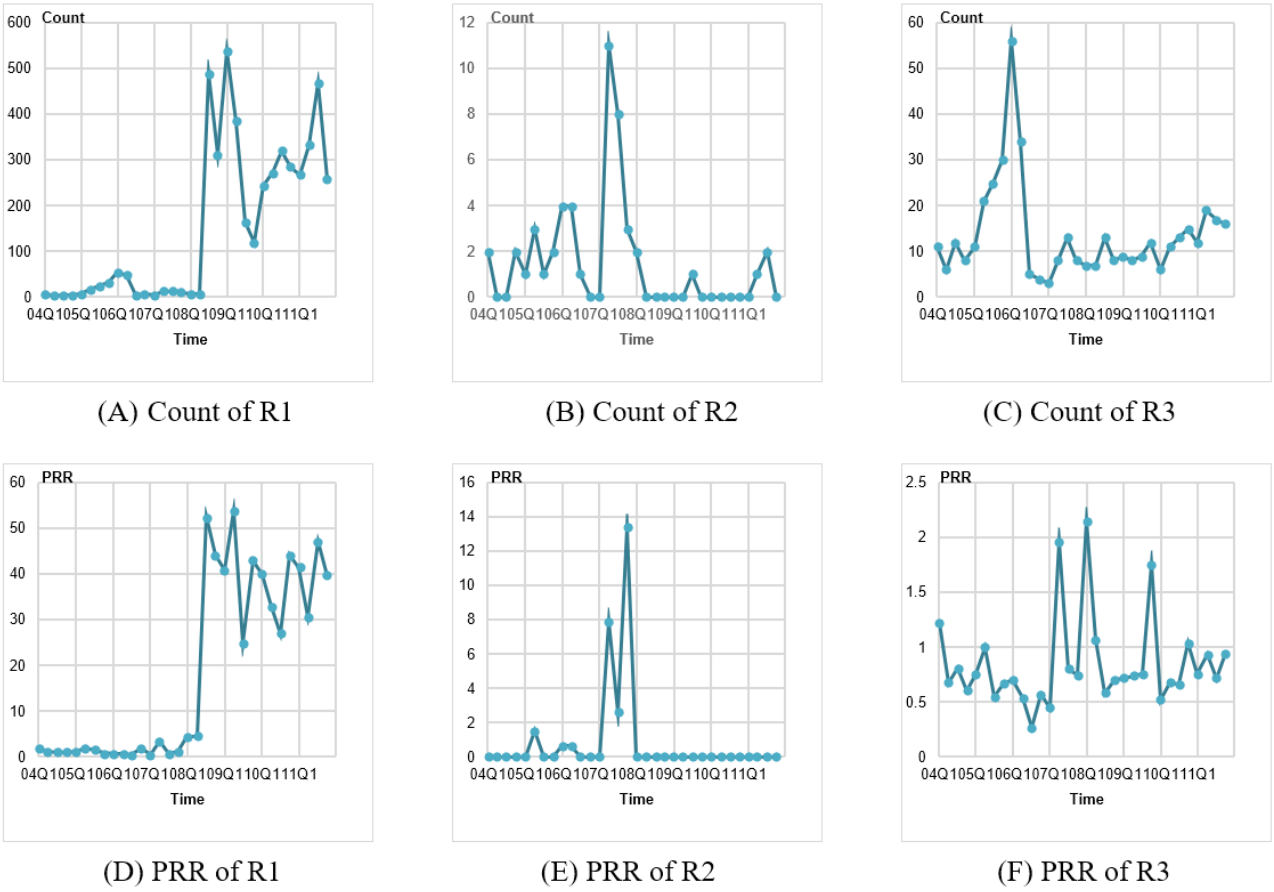
R3: Warfarin, Age>60 → Myocardial infarction

We calculated its occurrences, PRRs, and compared the values with the original values for each signal. We omit the results for uniform setting θ*=0.4 and level-wise setting because similar results were observed for uniform setting θ*=0.2 and frequency-based setting, respectively.

To highlight the impact of anonymization on rare events, we set PRR=0 when $a<3$, where a denotes the number of reports that satisfy the specific ADR rule. The threshold $a\geq3$ follows Evans et al [29], who investigated a group of newly marketed drugs and suggested that the minimum criteria for a signal are $a\geq3$ and PRR>2.

The original count and PRR of these 3 rules are shown in Figure 5. Rule R1 is a signal with an extremely high occurrence and significant strength, rule R2 is the one with the relatively small occurrence and medium strength, while R3 represents medium occurrence and relatively little strength.

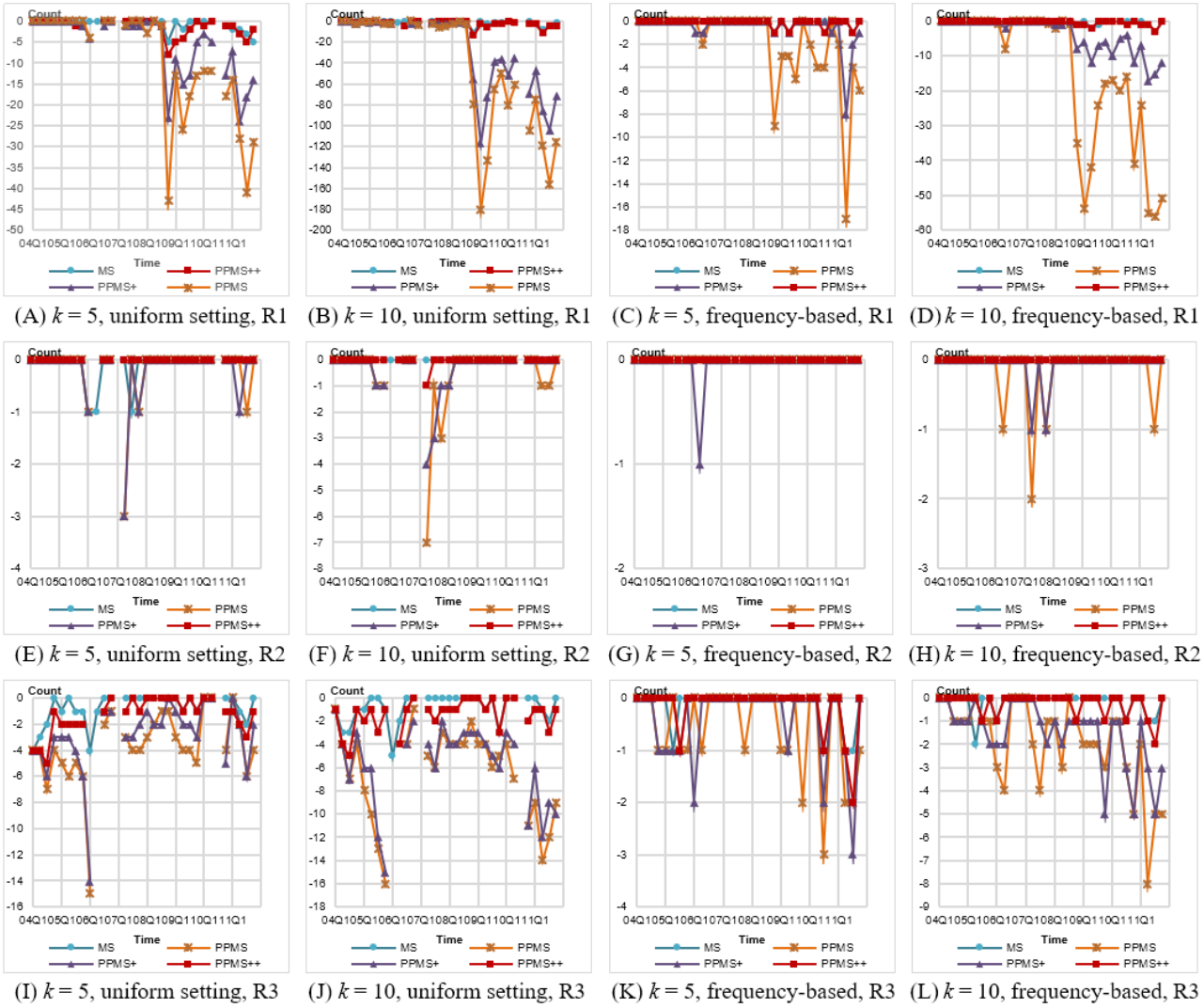
Figure 5. The original counts and proportional reporting ratios (PRRs) of rules R1, R2, and R3.



Signal Occurrence Variation

We first evaluated the variation of signal occurrence (count) caused by anonymization. The results are shown in Figure 6. Notice that there is no result for several quarters (eg, 2007Q1, 2010Q3) under the uniform setting. The reason is the same as that for information loss. Generally, the variation yielded by frequency-based setting is much less than that by uniform setting, and a larger k causes more missing counts. For signals with extremely high occurrence like R1, the variation can be substantial; for example, it reaches 180 for PPMS with k=10 and uniform confidence setting. In the same case, our PPMS^{++} exhibits outstanding performance, only causing variation of less than 10. We also note that some quarters are suffering significant count variation for rule R2 (Figure 6E-H). This is because the taxonomy of Gender is relatively flat, composed of only 2 levels. Once the gender of a report satisfying this rule is generalized, it will become “Any” and increase the missing count of this rule. For example, in Figure 6F, when k=10, 7 of 11 counts are missing in 2007Q2 for PPMS. In fact, when k=10, the ratio of reports with Gender=Any is at least 25% and 45% from 2010Q4 to 2011Q4 for PPMS^{+} and PPMS, respectively, which causes serious bias on the count of ADR rule. By contrast, as shown in Figure 6G and H, the frequency-based setting exhibits lower missing count. The overall situation shows that PPMS^{++} significantly outperforms PPMS and PPMS^{+}, and demonstrates comparable results with MS-Anonymization.

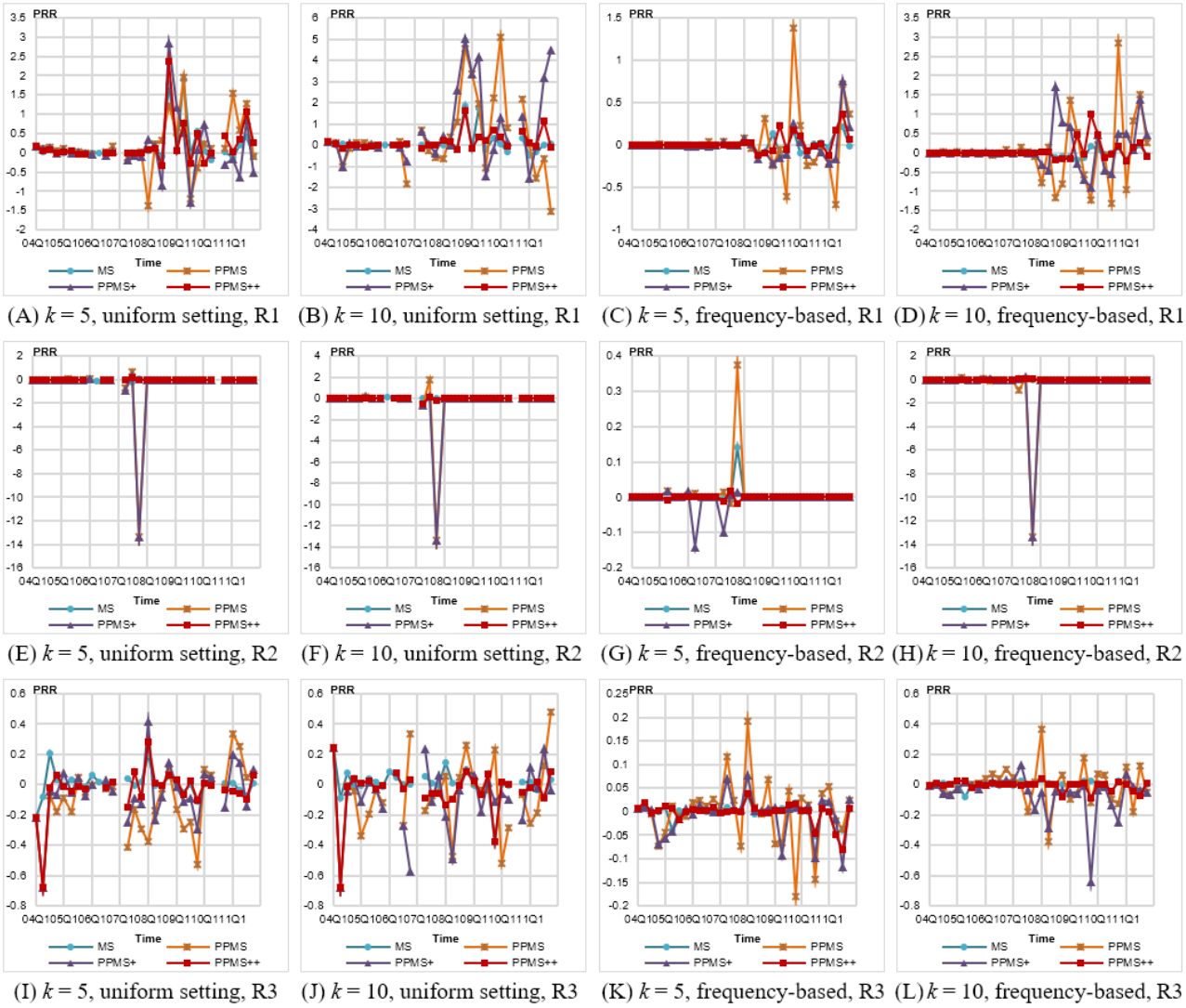
Figure 6. Variations in signal count for different anonymization methods under uniform and frequency-based settings of θ*. PPMS: periodical-publishing multisensitive.



Signal Strength Variation

Figure 7 shows the results on the PRR difference. Similar to that observed for occurrence variation, the frequency-based setting yields more negligible PRR difference than that by uniform setting. For rule R1 with enormous strength, the PRR variation is significantly higher than those for rules R2 and R3. The variations caused by PPMS and PPMS^{+} fluctuate seriously, sometimes much higher, reaching 5 for k=10 and uniform setting of θ*; PPMS^{++} exhibits relatively small variation under the same situation. For rule R2 with attributes of flat taxonomy, we observe a similar phenomenon. Specifically, a sharply significant variation, reaching –14 (Figure 7E, F, and H), is observed in 2007Q4 for PPMS and PPMS^{+}. This is because the a value for computing PRR is less than 3. We observe that the original count of this rule in 2007Q4 (Figure 5B) is 3 and its original PRR (Figure 5E) is 13.39. This means that this rule is a rare event with high strength. Any missing count of this rule causes value a to be less than 3 and the PRR will become 0, invalidating this rule. This situation demonstrates the impact of generalization on rare but significant ADR rule, especially for attributes with shallow generalization levels such as Gender, which will hinder or delay the discovery of ADR signals.

Figure 7. Variations in signal strength (proportional reporting ratio [PRR]) for different anonymization methods under uniform and frequency-based settings of θ*. PPMS: periodical-publishing multisensitive.



Discussion

Principal Results

In this paper, we have introduced the periodical publishing scenario usually adopted for publishing SRS data. We have presented 3 kinds of attacks, *BFL*-attacks, which exploit the CaseID of records to link the same cases in the series of releases to crack the anonymization by excluding the nontargets to improve the confidence to hit the record target or the sensitive value.

To prevent the record and attribute disclosure caused by *BFL*-attacks, we have presented a new model called PPMS(k, θ*)-bounding. We have also proposed an algorithm called PPMS-Anonymization to anonymize the raw SRS data set achieving the privacy requirement of PPMS(k, θ*)-bounding. Two enhancements of PPMS-Anonymization, PPMS^{+}-Anonymization and PPMS^{++}-Anonymization, have also been presented.

To evaluate the performance of our method, we conducted several experiments with different settings on privacy threshold, from 3 various aspects of evaluation, including information loss, PR, and bias on signal strength. The results showed that our proposed anonymization method, especially PPMS^{++}-Anonymization, can effectively prevent *BFL*-attacks caused by follow-up cases across a series of SRS data sets, guarantee the privacy requirement with controlled loss of data utility, and maintain the usability of anonymized SRS data set for ADR detection, especially for frequency-based threshold setting and level-wise setting.

Limitations

Fostering the development of new detection methods and early discovery of suspected ADR signals is the main driving force for many organizations such as the US FDA to release their SRS data sets to the public. By contrast, evaluating each individual case safety report (ICSR) is necessary for investigating hypothetical signals generated from the SRS data. Unfortunately, due to national privacy regulations such as the Health Insurance Portability and Accountability Act (HIPPA) Privacy Rule [30], some specified individual identifiers and narrative were removed from the published FAERS data (following the safe harbor method in Section 164.514 [30]). A recent work [31] showed that the absence of personal details would significantly affect the assessment of each ICSR. In this

context, the published SRS data alone cannot fulfill the purpose of ICSR evaluation. We endeavor to develop an effective privacy protection method for the partially deidentified SRS data (eg, FAERS) without sacrificing the data utility for aggregative disproportionality analysis of suspected ADR signals. How to protect the sharing and access of raw SRS data containing all individually identifiable health information is beyond the scope of this study. Instead, the SRS data organization should provide advanced security schemes, including technical or nontechnical [32], to ensure the confidentiality, integrity, and availability of the protected health information for authorized users, as enforced by the HIPPA Security Rule [33], which requires a good threat analysis modeling [34] before the system design.

Comparison With Prior Work

This paper is an extended version of our paper presented at IEEE *ICDE'17* [35]. Some new material has been added to clarify the design of the proposed PPMS-Anonymization and its improvement (PPMS+-Anonymization), including the design of the function Generalization (Multimedia Appendix 6), Multimedia Appendix 7, and Figure 1. A significantly more efficient version, PPMS++-Anonymization, is proposed. A new way of confidence threshold setting, level-wise setting, was evaluated. Additional more ADR signals were inspected. All experiments were reconducted to include the new version (PPMS++-Anonymization). Overall, PPMS++-Anonymization ensures zero PR on record and attribute linkage, while exhibits 51%-78% and 59%-82% improvements on information loss over PPMS+-Anonymization and PPMS-Anonymization, respectively, and significantly reduces the bias of ADR signal. For example, under the frequency setting, the maximum count bias and PRR bias were reduced from 56 to 3 and 13.4 to 0.1, respectively.

Based on our work [35], Huang et al [36] proposed 2 new attacks, *MD*-attack (Medicine Discontinuation attack) and *SS*-attack (Substantial Symptom attack). *MD*-attack assumes the attacker knew when the target stopped his/her treatment, that is, the quarter in which the target's follow-up record discontinues, while *SS*-attack regards a *QID* group with a substantial amount of adverse reactions risky. Both types of attacks, however, suffer some actuality problems. First, the authors overlooked the phenomenon that an individual's follow-up records may discontinue for some quarters and reappear in the next quarter. This life span discontinuity of follow-up cases is unpredictable and will thwart the justness of *MD*-attack and the anonymization algorithm. The problem for *SS*-attack is whether knowing someone having many adverse reactions does cause a privacy breach, which needs more convincing evidence. Besides, *SS*-attack is not related to periodical releases of SRS data.

Conclusions

In summary, our PPMS(k, θ*)-bounding and PPMS-Anonymization can anonymize SRS data sets in the periodical data publishing scenario, preventing the series of releases from the disclosure of sensitive personal information caused by *BFL*-attacks.

The *BFL*-attacks caused by the existence of CaseID in SRS data is not a particular case in health data. Other types of medical data contain similar features, for example, electronic health records, a digital version of a patient's paper chart composed of more private information than SRS data. As far as we know, it contains an attribute called patient ID which is similar to CaseID and so may be vulnerable to *BFL*-attacks. We will study this shortly. Some more challenging extensions of this topic include the study of incremental anonymization of data sets published in a cloud environment [37,38] and handling a large amount of missing values in SRS data [39]. Recently, the emerging differential privacy [40-42] has been widely recognized as a more rigorous privacy protection method [43]. Our recent work [44] on integrating differential privacy to anonymize a single release of SRS data has shown promising results. We are currently synergizing the differential privacy to our PPMS(k, θ*)-bounding to yield a better protection scheme.

Acknowledgments

This work was supported by the Ministry of Science and Technology of Taiwan under grant no. MOST103-2221-E-390-022.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A summary of privacy models for incremental data publishing.
[PDF File (Adobe PDF File), 59 KB - medinform_v9i10e28752_app1.pdf]

Multimedia Appendix 2
Proof of Theorem 1.
[PDF File (Adobe PDF File), 98 KB - medinform_v9i10e28752_app2.pdf]

Multimedia Appendix 3
Proof of Lemma 2.
[PDF File (Adobe PDF File), 75 KB - medinform_v9i10e28752_app3.pdf]

Multimedia Appendix 4
PPMS-Anonymization.
[PDF File (Adobe PDF File), 107 KB - medinform_v9i10e28752_app4.pdf]

Multimedia Appendix 5
Procedure Grouping.
[PDF File (Adobe PDF File), 101 KB - medinform_v9i10e28752_app5.pdf]

Multimedia Appendix 6
Function Generalization.
[PDF File (Adobe PDF File), 90 KB - medinform_v9i10e28752_app6.pdf]

Multimedia Appendix 7
Modification of PPMS-Anonymization to realize PPMS+-Anonymization.
[PDF File (Adobe PDF File), 93 KB - medinform_v9i10e28752_app7.pdf]

Multimedia Appendix 8
The taxonomy tree of Age.
[PDF File (Adobe PDF File), 82 KB - medinform_v9i10e28752_app8.pdf]

Multimedia Appendix 9
Modification of PPMS-Anonymization to realize PPMS++-Anonymization.
[PDF File (Adobe PDF File), 83 KB - medinform_v9i10e28752_app9.pdf]

Multimedia Appendix 10
Description of proportional reporting ratio.
[PDF File (Adobe PDF File), 30 KB - medinform_v9i10e28752_app10.pdf]

References

1. FDA Adverse Event Reporting System (FAERS). URL: https://open.fda.gov/data/faers/ [accessed 2017-04-30]
2. The Yellow Card Scheme. URL: http://yellowcard.mhra.gov.uk [accessed 2015-08-10]
3. MedEffect Canada. URL: https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada.html [accessed 2015-05-10]
4. Fung BCM, Wang K, Chen R, Yu PS, Mehta B. Privacy-preserving data publishing. ACM Comput. Surv 2010 Jun 01;42(4):1-53. [doi: 10.1145/1749603.1749605]
5. El Emam K, Dankar FK, Neisa A, Jonker E. Evaluating the risk of patient re-identification from adverse drug event reports. BMC Med Inform Decis Mak 2013 Oct 05;13:114 [FREE Full text] [doi: 10.1186/1472-6947-13-114] [Medline: 24094134]
6. Sweeney L. k-anonymity: a model for protecting privacy. Int. J. Unc. Fuzz. Knowl. Based Syst 2012 May 02;10(05):557-570. [doi: 10.1142/S0218488502001648]
7. Lin WY, Yang DC. On privacy-preserving publishing of spontaneous ADE reporting data. In: Proceedings of 2013 IEEE International Conference on Bioinformatics and Biomedicine. 2013 Dec 18 Presented at: 2013 IEEE International Conference on Bioinformatics and Biomedicine; December 18-21, 2013; Shanghai, China p. 51-53. [doi: 10.1109/BIBM.2013.6732760]
8. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 2007 Mar 01;1(1):3-es. [doi: 10.1145/1217299.1217302]
9. Lin WY, Yang DC, Wang JT. Privacy preserving data anonymization of spontaneous ADE reporting system dataset. BMC Med Inform Decis Mak 2016 Jul 18;16 (Suppl 1):58 [FREE Full text] [doi: 10.1186/s12911-016-0293-4] [Medline: 27454754]
10. Lin WY, Lo CF. Co-training and ensemble based duplicate detection in adverse drug event reporting systems. In: Proceedings of 2013 IEEE International Conference on Bioinformatics and Biomedicine. 2013 Dec 18 Presented at: 2013 IEEE International Conference on Bioinformatics and Biomedicine; December 18-21, 2013; Shanghai, China p. 7-8. [doi: 10.1109/bibm.2013.6732591]
11. Tregunno PM, Fink DB, Fernandez-Fernandez C, Lázaro-Bengoa E, Norén GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Saf 2014 Mar 14;37(4):249-258. [doi: 10.1007/s40264-014-0146-y] [Medline: 24627310]
12. Kreimeyer K, Menschik D, Winiecki S, Paul W, Barash F, Woo EJ, et al. Using probabilistic record linkage of structured and unstructured dData to identify duplicate cases in spontaneous adverse event reporting systems. Drug Saf 2017 Mar 14;40(7):571-582. [doi: 10.1007/s40264-017-0523-4] [Medline: 28293864]

13. Byun JW, Sohn Y, Bertino E, Li N. Secure anonymization for incremental data sets. 2006 Presented at: The 3rd VLDB Workshop on Secure Data Management; September 10-11, 2006; Seoul, Korea p. 48-63. [doi: 10.1007/11844662_4]
14. Byun JW, Li T, Bertino E, Li N, Sohn Y. Privacy-preserving incremental data dissemination. JCS 2009 Mar 16;17(1):43-68. [doi: 10.3233/jcs-2009-0316]
15. Pei J, Xu J, Wang Z, Wang W, Wang K. Maintaining k-anonymity against incremental updates. In: Proceedings of the 19th International Conference on Scientific and Statistical Database Management. 2007 Jul 09 Presented at: The 19th International Conference on Scientific and Statistical Database Management; July 9-11, 2007; Banff, Canada p. 5-14. [doi: 10.1109/ssdbm.2007.16]
16. Fung BCM, Wang K, Fu AWC, Pei J. Anonymity for continuous data publishing. In: Proceedings of the 11th International Conference on Extending Database Technology. 2008 Mar 25 Presented at: The 11th International Conference on Extending Database Technology; March 25-29, 2008; Nantes, France p. 264-275. [doi: 10.1145/1353343.1353378]
17. Xiao X, Tao Y. M-invariance: towards privacy preserving re-publication of dynamic data sets. In: Proceedings of 2007 ACM SIGMOD International Conference on Management of Data. 2007 Jun 11 Presented at: The 2007 ACM SIGMOD International Conference on Management of Data; June 11-14, 2007; Beijing, China p. 689-700. [doi: 10.1145/1247480.1247556]
18. Bu Y, Fu AWC, Wong RCW, Chen L, Li J. Privacy preserving serial data publishing by role composition. Proc. VLDB Endow 2008 Aug;1(1):845-856. [doi: 10.14778/1453856.1453948]
19. Li F, Zhou S. Challenging more updates: towards anonymous re-publication of fully dynamic data sets. arXiv. 2008 Jun 28. URL: https://arxiv.org/abs/0806.4703 [accessed 2021-05-22]
20. Anjum A, Raschia G, Gelgon M, Khan A, Malik SUR, Ahmad N, et al. τ -safety: A privacy model for sequential publication with arbitrary updates. Computers & Security 2017 May;66:20-39. [doi: 10.1016/j.cose.2016.12.014]
21. He Y, Barman S, Naughton JF. Preventing equivalence attacks in updated, anonymized data. In: Proceedings of the 27th IEEE International Conference on Data Engineering. 2011 Apr Presented at: The 27th IEEE International Conference on Data Engineering; April 11-16, 2011; Hannover, Germany p. 529-540. [doi: 10.1109/icde.2011.5767924]
22. Bewong M, Liu J, Liu L, Li J. Privacy preserving serial publication of transactional data. Information Systems 2019 May;82:53-70. [doi: 10.1016/j.is.2019.01.001]
23. Wang K, Fung BCM. Anonymizing sequential release. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006 Aug Presented at: The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 20-23, 2006; New York, NY p. 414-423. [doi: 10.1145/1150402.1150449]
24. Shmueli E, Tassa T, Wasserstein R, Shapira B, Rokach L. Limiting disclosure of sensitive data in sequential releases of databases. Information Sciences 2012 May;191:98-127. [doi: 10.1016/j.ins.2011.12.020]
25. Shmueli E, Tassa T. Privacy by diversity in sequential releases of databases. Information Sciences 2015 Mar;298:344-372. [doi: 10.1016/j.ins.2014.11.005]
26. Byun JW, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. In: Proceedings of the 12th International Conference on Database Systems for Advanced Applications. 2007 Apr Presented at: The 12th International Conference on Database Systems for Advanced Applications; April 9-12, 2007; Bangkok, Thailand p. 188-200. [doi: 10.1007/978-3-540-71703-4_18]
27. Medical Subject Headings (MeSH). URL: http://www.ncbi.nlm.nih.gov/mesh/ [accessed 2017-03-10]
28. MedWatch. URL: http://www.fda.gov/Safety/MedWatch/ [accessed 2015-08-10]
29. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001 Dec 10;10(6):483-486. [doi: 10.1002/pds.677] [Medline: 11828828]
30. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Office for Civil Rights. 2012 Nov. URL: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf [accessed 2021-06-07]
31. Marwitz K, Jones SC, Kortepeter CM, Dal Pan GJ, Muñoz MA. An evaluation of postmarketing reports with an outcome of death in the US FDA adverse event reporting System. Drug Saf 2020 May;43(5):457-465. [doi: 10.1007/s40264-020-00908-5] [Medline: 31981082]
32. Scheibner J, Raisaro JL, Troncoso-Pastoriza JR, Ienca M, Fellay J, Vayena E, et al. Revolutionizing medical data sharing using advanced privacy-enhancing technologies: technical, legal, and ethical synthesis. J Med Internet Res 2021 Feb 25;23(2):e25120 [FREE Full text] [doi: 10.2196/25120] [Medline: 33629963]
33. Summary of the HIPAA Security Rule. Office for Civil Rights. 2013 Jul. URL: https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html [accessed 2021-06-17]
34. Shevchenko N, Chick TA, O'Riordan P, Scanlon TP, Woody C. Threat Modeling: A Summary of Available Methods. URL: https://resources.sei.cmu.edu/asset_files/WhitePaper/2018_019_001_524597.pdf [accessed 2021-07-21]
35. Wang JT, Lin WY. Privacy preserving anonymity for periodical SRS data publishing. In: Proceedings of the 33rd IEEE International Conference on Data Engineering. 2017 Apr Presented at: The 33rd IEEE International Conference on Data Engineering; April 19-22, 2017; San Diego, CA p. 1344-1355. [doi: 10.1109/icde.2017.176]

36. Huang W, Yi T, Zhu H, Shang W, Lin W. Improved privacy preserving method for periodical SRS publishing. PLoS One 2021 Apr 22;16(4):e0250457 [FREE Full text] [doi: 10.1371/journal.pone.0250457] [Medline: 33886662]
37. Aldeen YAAS, Salleh M, Aljeroudi Y. An innovative privacy preserving technique for incremental datasets on cloud computing. J Biomed Inform 2016 Aug;62:107-116 [FREE Full text] [doi: 10.1016/j.jbi.2016.06.011] [Medline: 27369566]
38. Jeon S, Seo J, Kim S, Lee J, Kim JH, Sohn JW, et al. Proposal and assessment of a de-identification strategy to enhance anonymity of the observational medical outcomes partnership common data model (OMOP-CDM) in a public cloud-computing environment: anonymization of medical data using privacy models. J Med Internet Res 2020 Nov 26;22(11):e19597 [FREE Full text] [doi: 10.2196/19597] [Medline: 33177037]
39. Hsiao MH, Lin WY, Hsu KY, Shen ZX. On anonymizing medical microdata with large-scale missing values - A case study with the FAERS dataset. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society. 2019 Jul Presented at: The 41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society; July 23–27, 2019; Berlin, Germany p. 6505-6508. [doi: 10.1109/EMBC.2019.8857025]
40. Dwork C. Differential privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming. 2006 Jul Presented at: The 33rd International Conference on Automata, Languages and Programming; July 10-14, 2006; Venice, Italy p. 1-12. [doi: 10.1007/11787006_1]
41. Liu F. Generalized Gaussian Mechanism for Differential Privacy. IEEE Trans. Knowl. Data Eng 2019 Apr 1;31(4):747-756. [doi: 10.1109/tkde.2018.2845388]
42. Wang D, Xu Z. Impact of inaccurate data on differential privacy. Computers & Security 2019 May;82:68-79. [doi: 10.1016/j.cose.2018.12.007]
43. Desfontaines D, Pejó B. SoK: Differential privacies. Proceedings on Privacy Enhancing Technologies 2020 May;2:288-313. [doi: 10.2478/popets-2020-0028]
44. Lin WY, Shen ZX. Embracing differential privacy for anonymizing spontaneous ADE reporting data. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine. 2020 Presented at: The 2020 IEEE International Conference on Bioinformatics and Biomedicine; December 16-19, 2020; Seoul, Korea p. 2015-2022. [doi: 10.1109/bibm49941.2020.9313578]

Abbreviations

ADE: adverse drug event
ADR: adverse drug reaction
DIG: dangerous identity group
DIR: dangerous identity ratio
DSG: dangerous sensitivity group
DSR: dangerous sensitivity ratio
FAERS: FDA Adverse Event Reporting System
FDA: Food and Drug Administration
HIPPA: Health Insurance Portability and Accountability Act
MedDRA: Medical Dictionary for Regulatory Activities
MHRA: UK Medicines and Healthcare products Regulatory Agency
NIL: normalized information loss
PPDP: privacy-preserving data publishing
PPMS: periodical-publishing multisensitive
PRR: proportional reporting ratio
QID: quasi-identifier
SA: sensitive attribute
SRS: spontaneous reporting system

Edited by G Eysenbach; submitted 13.03.21; peer-reviewed by P Natsavias; comments to author 25.04.21; revised version received 30.07.21; accepted 02.08.21; published 28.10.21.

Please cite as:
Wang JT, Lin WY
Privacy-Preserving Anonymity for Periodical Releases of Spontaneous Adverse Drug Event Reporting Data: Algorithm Development and Validation
JMIR Med Inform 2021;9(10):e28752
URL: https://medinform.jmir.org/2021/10/e28752
doi:10.2196/28752
PMID:34709197

Original Paper

Categorizing Vaccine Confidence With a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse

Per E Kummervold[1], PhD; Sam Martin[2,3,4,5], LLB, MSc, PhD; Sara Dada[4,6], MSc; Eliz Kilich[4], BMBCh; Chermain Denny[7], BSc; Pauline Paterson[4,8], PhD; Heidi J Larson[4,8,9,10], PhD

[1]Vaccine Research Department, FISABIO-Public Health, Valencia, Spain

[2]Centre for Clinical Vaccinology and Tropical Medicine, University of Oxford, Oxford, United Kingdom

[3]Rapid Research Evaluation and Appraisal Lab, Departament of Targeted Intervention, University College London, London, United Kingdom

[4]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, United Kingdom

[5]Ethox Centre, Nuffield Department of Population Health, Big Data Institute, University of Oxford, Oxford, United Kingdom

[6]UCD Centre for Interdisciplinary Research, Education and Innovation in Health Systems, School of Nursing, Midwifery and Health Systems, University College Dublin, Dublin, Ireland

[7]Faculty of Science, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[8]NIHR Health Protection Research Unit, London, United Kingdom

[9]Institute of Health Metrics and Evaluation, University of Washington, Seattle, WA, United States

[10]Chatham House Centre on Global Health Security, The Royal Institute of International Affairs, London, United Kingdom

Corresponding Author:
Per E Kummervold, PhD
Vaccine Research Department
FISABIO-Public Health
Avda. de Catalunya, 21
Valencia, 46020
Spain
Phone: 34 41435795
Email: per@capia.no

Abstract

Background: Social media has become an established platform for individuals to discuss and debate various subjects, including vaccination. With growing conversations on the web and less than desired maternal vaccination uptake rates, these conversations could provide useful insights to inform future interventions. However, owing to the volume of web-based posts, manual annotation and analysis are difficult and time consuming. Automated processes for this type of analysis, such as natural language processing, have faced challenges in extracting complex stances such as attitudes toward vaccination from large amounts of text.

Objective: The aim of this study is to build upon recent advances in transposer-based machine learning methods and test whether transformer-based machine learning could be used as a tool to assess the stance expressed in social media posts toward vaccination during pregnancy.

Methods: A total of 16,604 tweets posted between November 1, 2018, and April 30, 2019, were selected using keyword searches related to maternal vaccination. After excluding irrelevant tweets, the remaining tweets were coded by 3 individual researchers into the categories *Promotional*, *Discouraging*, *Ambiguous*, and *Neutral or No Stance*. After creating a final data set of 2722 unique tweets, multiple machine learning techniques were trained on a part of this data set and then tested and compared with the human annotators.

Results: We found the accuracy of the machine learning techniques to be 81.8% (*F* score=0.78) compared with the agreed score among the 3 annotators. For comparison, the accuracies of the individual annotators compared with the final score were 83.3%, 77.9%, and 77.5%.

Conclusions: This study demonstrates that we are able to achieve close to the same accuracy in categorizing tweets using our machine learning models as could be expected from a single human coder. The potential to use this automated process, which is

reliable and accurate, could free valuable time and resources for conducting this analysis, in addition to informing potentially effective and necessary interventions.

(JMIR Med Inform 2021;9(10):e29584) doi:10.2196/29584

KEYWORDS

computer science; information technology; public health; health humanities; vaccines; machine learning

Introduction

Background

Although individuals have been found to share different thoughts, questions, and concerns about vaccines on social media [1], studies of the vaccine discourse on social media [2] indicate that concerns, and indeed the sharing of misinformation in particular, are amplified [3]. What is of concern is the number of imprecise and inaccurate articles available with regard to vaccinations.

Multiple studies have been conducted to monitor vaccination discussions on social media [4-6]. Addressing misunderstandings and inaccuracies as early as possible is vital to making sound vaccine policies. However, there is currently insufficient research on how to effectively categorize the nuances in the perceptions of sentiment toward vaccines in the large volume of vaccine data shared daily on social media. Being able to monitor and understand the spread and traction of misinformation in social media on a larger global scale is key to mitigating the negative impact of such information.

Although the data retrieved from social and news media might not be representative of the entire population, they provide a snapshot of discussions and thoughts, and the trends observed here are still thought to be of vital importance to understanding emerging issues of concern as well as the link between misinformation on news and social media platforms and the effect of this misinformation on vaccination confidence and uptake. To detect such trends, however, we need an in-depth understanding of the content of these messages. Although qualitative methods might provide this insight, the sheer volume of news and social media content makes it difficult to apply these methods to conversations among entire populations over time. Machine learning and natural language processing (NLP) have the potential to handle huge amounts of information. However, concerns over accuracy, especially when dealing with the complexity of the language used to express opinions about vaccines, have prevented these methods from being very effective.

Sentiment analysis in machine learning refers to the process of automatically determining whether the author of a piece of text is in favor of, against, or neutral toward the subject of the statement. This is slightly different from stance detection, which involves automatically determining the author's attitude toward a proposition or target [7]. Although sentiment analysis can look only at the tone of a particular statement, stance detection often refers to a target outside of the particular statement.

The author of a tweet could express a positive attitude toward vaccination by expressing negativity toward people opposing vaccines (for instance, so-called *antivaxxers*). This double negation would then be interpreted as *positive* or *promotional*. This could be referred to as the *author's sentiment toward vaccination*, but because *sentiment* is often used for referring to the *sentiment of the statement*, we find it less confusing to refer to this as the *author's stance* toward vaccination. This distinction is particularly important when studying an issue as complex as vaccination because many texts often express strong opinions about vaccination without addressing vaccines directly. The distinction can be illustrated using the examples presented in Table 1.

Historically, NLP has often focused on ordinary sentiment analysis. This is technically a much easier task, but it is less useful from a sociological point of view. In contrast to *sentiment*, a person's *stance* toward a target can be expressed using either negative or positive language. People could, for instance, switch from opposing *abortion* to promoting *prolife* without changing their basic stance. In a sociological analysis, we would usually be more interested in the stance that people have toward a topic or target than the sentiment expressed in a particular statement.

Table 1. Difference between sentiment and stance.

Text	Sentiment (subject)	Stance (target)
Vaccines save lives	Positive (vaccines)	Positive (vaccines)
Antivaxxers kill people with their misinformation	Negative (antivaxxers)	Positive (vaccines)
Trust your doctor's knowledge regarding vaccines	Positive (physician's knowledge)	Positive (vaccines)
Antivaxxers tell the real truth about vaccines	Positive (antivaxxers)	Negative (vaccines)

Objective

The aim of this study is to look more deeply into attitudes toward maternal vaccination and how well the task of detecting stance in tweets can be accomplished by using multiple machine learning methods. We attempt to quantify how accurately such tweets can be categorized by trained annotators and how this compares with newer machine learning methods.

Methods

Overview

This research collected 16,605 Twitter messages (tweets) published over 6 months between November 1, 2018, and April 30, 2019, from Meltwater [8], a media intelligence system. This data set was collected and coded to complement a larger research study on sentiments and experiences around maternal vaccination across 15 countries (Australia, Brazil, Canada, France, Germany, India, Italy, Mexico, Panama, South Africa, South Korea, Spain, Taiwan, the United Kingdom, and the United States). Non-English tweets were translated into English using Google Translate script (Alphabet Inc). Multimedia Appendix 1 includes the search queries used in this study. Before annotating, all usernames and links were replaced by a common tag. This served two purposes: it preserved anonymity, and it limited potential bias based on the coder's interpretation of the username. The target of the analysis should be to decipher what the text is actually telling the reader about the writer's stance toward vaccination.

In this study, *maternal vaccination* typically refers to the vaccines that are recommended by health authorities for pregnant women.

Individual tweets were manually coded into stance categories (Textbox 1). Stance was categorized across four sentiments toward maternal vaccines: *Promotional* (in favor of maternal vaccines), *Ambiguous* (uncertainty with mixed sentiment toward maternal vaccines), *Discouraging* (against maternal vaccines), and *No stance* (statements or facts about maternal vaccines that do not reveal the author's stance). Although it can be argued that some of the categories can be ordered, we treated them as nominal variables, not ordinal variables, in the analysis. Therefore, a tweet stating that pregnant women should take the tetanus vaccine but not the measles vaccine is considered a promotional post in favor of maternal vaccines because it encourages following the current health recommendations.

Textbox 1. Stance categorized across four sentiments toward maternal vaccines.

Stance categories and their definitions

- Promotional
  - Posts communicate public health benefits or safety of maternal vaccination.
  - Posts contain positive tones, supportive or encouraging toward maternal vaccination.
- Ambiguous
  - Posts contain indecision and uncertainty on the risks or benefits of maternal vaccination, or they are ambiguous.
  - Posts contain disapproving and approving information.
  - Posts describe risks of not vaccinating during pregnancy.
  - Posts refute claims that maternal vaccines are dangerous.
- Discouraging
  - Posts contain negative attitudes toward, or arguments against maternal vaccines.
  - Posts contain questions regarding effectiveness or safety or possibility of adverse reactions (eg, links to disability or autism).
  - Posts discourage the use of recommended maternal vaccines.
- Neutral or no stance
  - Posts contain no elements of uncertainty or promotional or negative content. These are often not sentiments expressed on the web but rather statements that are devoid of emotion. This category includes factual posts pointing to articles on maternal vaccines (eg, Study on effectiveness of maternal flu vaccine).

Cleaning the Data Set

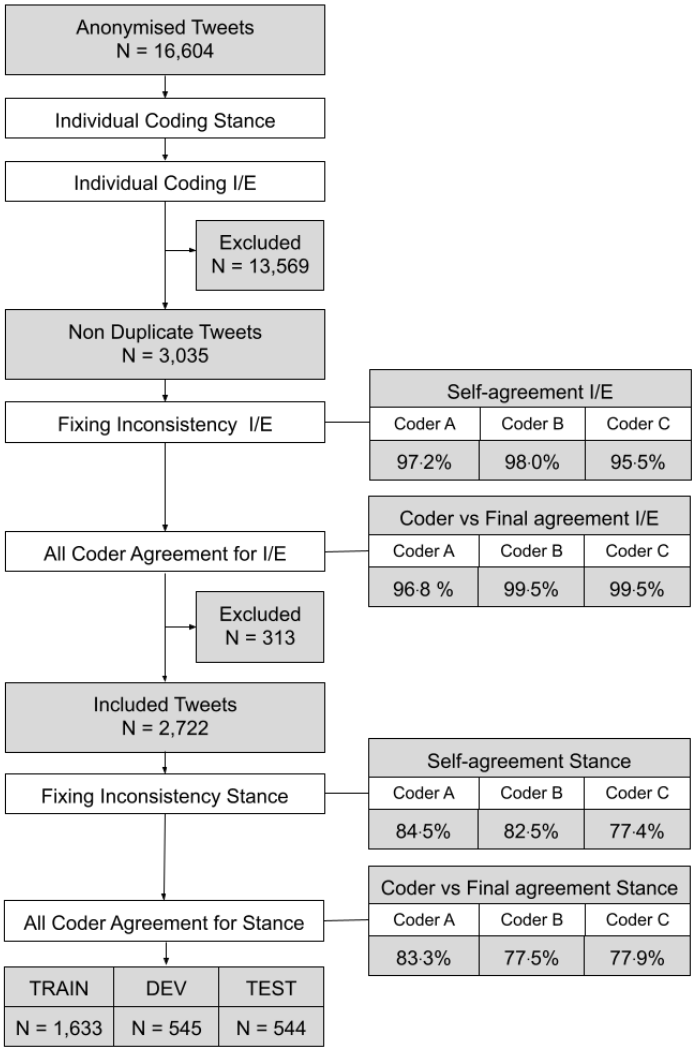
After the initial annotating, the data set was cleaned for duplicates and semiduplicates. Semiduplicates are tweets in which a few characters differ but the meaning is unchanged. A typical example is a retweeted post with the *RT:* prefix. Another example is a tweet suffixed (by a user or bot) with a few random characters to avoid being recognized (by Twitter detection algorithms) as a mass posting. To detect semiduplicates, we used a nonnormalized Levenshtein distance of less than 30 for tweets with more than 130 characters. For shorter tweets, the distance was scaled. The validity of the deduplication algorithm was qualitatively evaluated by the annotators. We were aiming for a *greedy* algorithm that identified too many semiduplicates rather than too few. Although this could slightly affect the size of the training set, it was considered to be of greater importance to prevent tweets that looked too similar from being included in both the training and test data sets. We have open sourced the Python code that we developed for cleaning and removing duplicates and made it available in our web-based GitHub repository [9].

As the stance of the tweet should be determined solely based on the content of the text, we deidentified usernames and URLs. Apart from cleaning usernames and URLs, the aim was to ensure that the input to the machine learning algorithm was exactly the

same as that presented to the human annotators. In this respect, this served to both preserve anonymity and limit potential bias arising from the annotator's interpretation of the usernames. This also helped fulfill ethical responsibilities as well as the European Union's General Data Protection Regulation guidelines. In all, 3 independent annotators screened all posts for inclusion, excluding posts that were not about the vaccines administered during pregnancy. The annotators also met to agree on the final posts included in the analysis [10].

Deduplication was conducted after the first round of annotating, and the annotators were then asked to recode any tweet for which they had provided inconsistent annotating. For example, there were instances where the same coder coded identical tweets inconsistently. From the tweets that appeared only twice in the material, we calculated a self-agreement score both for include or exclude and for stance. This was done to illustrate some of the potential challenges of manual annotating (Figure 1).

Figure 1. Screening and annotating procedure. DEV: development data set; I/E: include or exclude.



Bidirectional Encoder Representations From Transformers

The main model was based on the newest (May 2019) Whole Word Masking variant of Google's Bidirectional Encoder Representations from Transformers (BERT) [11]. When published in late 2018, the model demonstrated state-of-the-art results on 11 NLP tasks, including the Stanford Question Answering Dataset version 1.1 (The Stanford Natural Language Processing Group) [12]. BERT is a bidirectional, contextual encoder built on a network architecture called Transformer, which is based solely on attention mechanisms [13]. The main part of the training can be performed on unsupervised and unlabeled text corpora such as Wikipedia, and the pretrained weights [14] are trained solely on this general corpus.

We trained the model on a domain-specific corpus to expose the model to the vocabulary that is typically used in vaccination posts. We started creating domain-specific pretraining data by downloading 5.9 million tweets acquired by keyword searches related to vaccine and vaccination (Multimedia Appendix 2). The set was downloaded from Meltwater and preprocessed in the same way as the maternal vaccine tweets (ie, deduplication and username or link anonymization). The BERT architecture depends on carrying out unsupervised pretraining using a technique called Masked Language Modeling and Next Sentence Prediction. The latter method requires each text segment to have

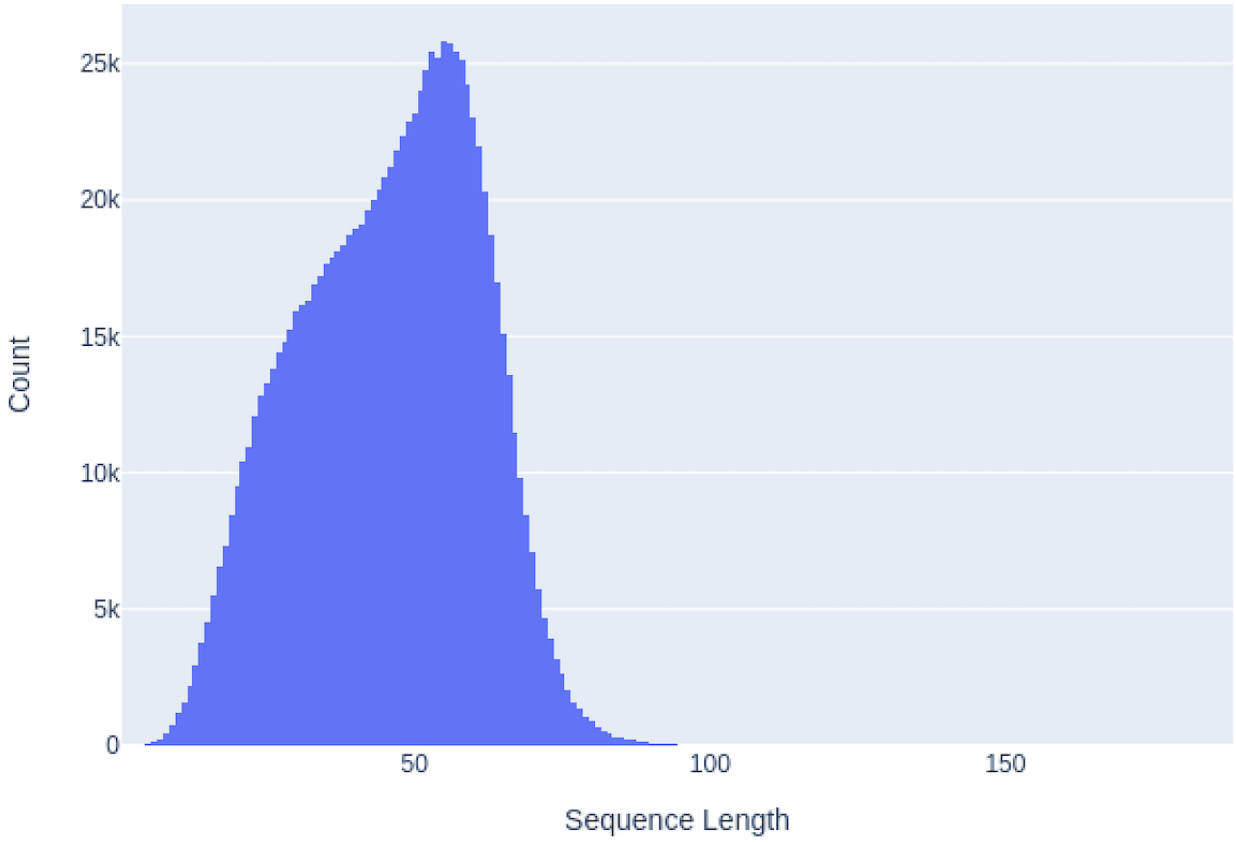
at least two sentences. Therefore, we filtered out all tweets that did not satisfy this criterion, reducing the data set from 1.6 million to 1.1 million tweets. A later study by Liu et al [15] pointed out that the Next Sentence Prediction task is not as effective as expected and that it might be beneficial to train the model only on the Masked Language Modeling task. As we have such a large number of short texts, this would have extended our data set. We refer to this data set as the vaccine-tweet data set.

We tokenized the tweets using the BERT vocabulary and limited the sequence length to 96 tokens. By limiting the sequence length, we were able to increase the batch size, which is known to have a positive effect on performance. Figure 2 shows the sequence length of the downloaded tweets, showing that this trimming would affect less than one in thousand tweets. Tweets longer than 96 tokens were manually examined, confirming that these were mainly repetitive sequences and that the trimming did not affect the meaning (eg, a statement followed by strings of varying lengths of repeated characters such as ..... or *????*). We have further addressed discourse distribution, labeling, and word balance in the word corpus in the study by Martin et al [10].

In addition, we acquired a data set with a total of 201,133 vaccine-related news articles from the Vaccine Confidence Project (London School of Hygiene & Tropical Medicine) media archive. The articles were collected by automated keyword searches from several sources, including Google News, HealthMap, and Meltwater. It is an extensive collection of vaccine-related articles in English from both news media and blogs. The search criteria have been developed over the years, which is why they have varied slightly, but they are very similar to the list presented in Multimedia Appendix 2. We refer to this data set as the vaccine news data set. We chose not to pretrain the model on a maternal vaccine–specific data set because we wanted the encoder representations to also be used on other vaccine-related topics. All pretraining was carried out using a learning rate of 2e-5, a batch size of 216, and a maximum sequence length of 96.

These domain-specific pretrained weights were the starting points for the classification of the maternal vaccination tweets. The manually classified maternal vaccination tweets were preprocessed in the same way as the tweets in the vaccine-tweet data set and then divided into training, development, and test data sets in the ratio 60:20:20 (N=1633:545:544).

Figure 2. Number of tokens in each tweet (count per million tweets).



Fine-tuning

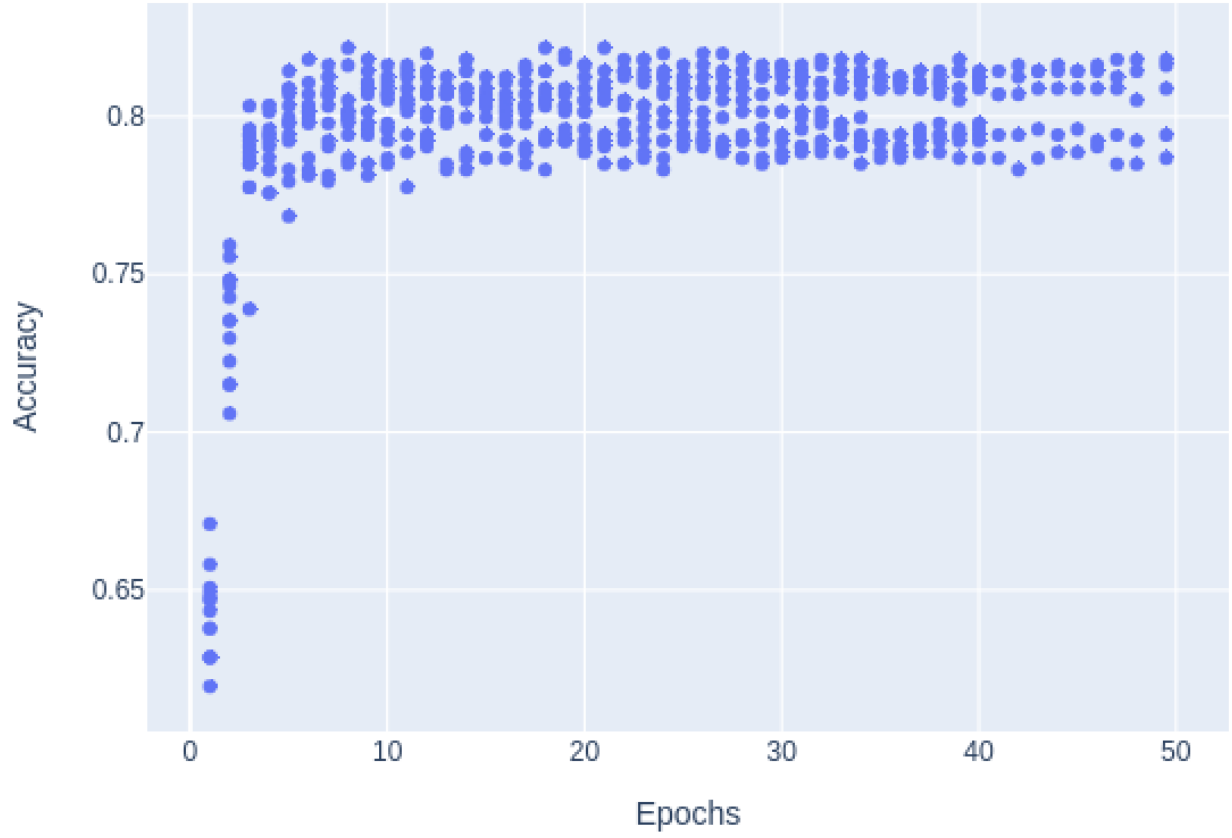
The pretraining of transposer models is a very slow process, but when these pretrained weights are determined, the final fine-tuning step is fast. To our knowledge, the best way of comparing the various pretrained models is by comparing their performances after fine-tuning. Figure 3 shows that the fine-tuning did not improve performance after 15 epochs but that there was considerable variance among the runs. For this reason, all pretrained models were evaluated using an average of 10 fine-tuned runs each at 15 epochs.

To obtain a baseline score for comparative machine learning models, various traditional and well-established networks were trained. The aim was to use well-established networks with known performance against standardized data sets for sentiment and stance analysis. The benchmark architectures, the neural network, and the long short-term memory (LSTM) networks with and without Global Vectors for Word Representation (GloVe; The Stanford Natural Language Processing Group) word embeddings were all taken from *Deep Learning With Python* by Chollet [16].

To verify that the neural network was able to solve other neural network tasks, we tested the network structures on one of the most basic NLP tasks: predicting positive and negative sentiments in IMDb (Amazon, Inc) movie reviews [17].

The final domain-specific pretraining and fine-tuning were carried out on a Cloud TPU (Tensor Processing Unit) v2-8 node with 8 cores and 64 gibibytes memory and an estimated performance of 180 teraflops (Google Cloud). Domain-specific pretraining was carried out for 2 weeks, but, as shown in Figure 3, there were no measurable improvements after a few days of running. Fine-tuning requires fewer computing resources and is usually completed in a few minutes on this platform.

Figure 3. Data evaluation of pretraining accuracy.



Results

Overview

In total, 3 annotators individually coded 2722 tweets. Of these 2722 tweets, 1559 (57.27%) were coded identically, with a Fleiss agreement score of κ=0.56. After meeting and discussing the tweets that they disagreed on, the annotators agreed on the annotating of all the remaining tweets. Although the annotators agreed on a final category for every tweet, they also reported that 6.83% (186/2722) of tweets "could be open to interpretation." Comparing the final agreed annotating after the discussions with the annotators' initial annotating, the accuracies of the individual annotators were 83.3%, 77.9%, and 77.5%. The accuracy of the machine learning model was also calculated with regard to the final agreed annotating.

One of the basic neural networks for NLP consists of two fully connected layers. For our data set, this only provided an accuracy of 43.7%. Thus, the network was not able to obtain a better result than simply predicting the overrepresented task *Promotional* for all data points. Adding pretrained GloVe word embeddings to this structure resulted in a slightly better performance, with a maximum accuracy of 55.5% on the test data set. However, both approaches overtrained after just a few epochs with data sets of this size.

To evaluate the reason for this low accuracy, we tested the same network on the IMDb data set, setting the same number of training examples (N=1633). In this case, the network achieved an accuracy of more than 80% even without the GloVe word embeddings, showing that the low accuracy was related to the difficulty of the maternal vaccine categorization.

The modern LSTM model is a recurrent neural network with a memory module. This model architecture was considered state-of-the-art a couple of years ago. We were able to obtain an accuracy of 63.1% here and can improve this to 65.5% by adding pretrained GloVe word embeddings.

Main Research Target

Our main research target was to investigate whether state-of-the-art NLP models could be improved by using the BERT architecture. Using the original pretrained weights, we achieved an average accuracy of 76.7% when fine-tuning for 15 epochs.

Starting from the original weights, the model weights were pretrained on the larger vaccine news data set for 1 million epochs. At various checkpoints (0 E, 250,000 E, 500,000 E, 750,000 E, and 1,000,000 E), the model was forked and then trained on the smaller and more specific vaccine-tweet data set.

At each of the checkpoints, the network was fine-tuned on the manually labeled tweets for 15 epochs, and the average of 10 runs is shown in Figure 4. Using pretraining on domain-specific content, the accuracy reached a peak of approximately 79% when training only on the vaccine news data set. However, by training first on the vaccine news data set for 250,000 epochs and then on the vaccine news data set for an additional 200,000 epochs, we were able to obtain an accuracy of 81.8%. For a comparison of accuracies, see Table 2.

Figure 4. Average of pretraining accuracy.

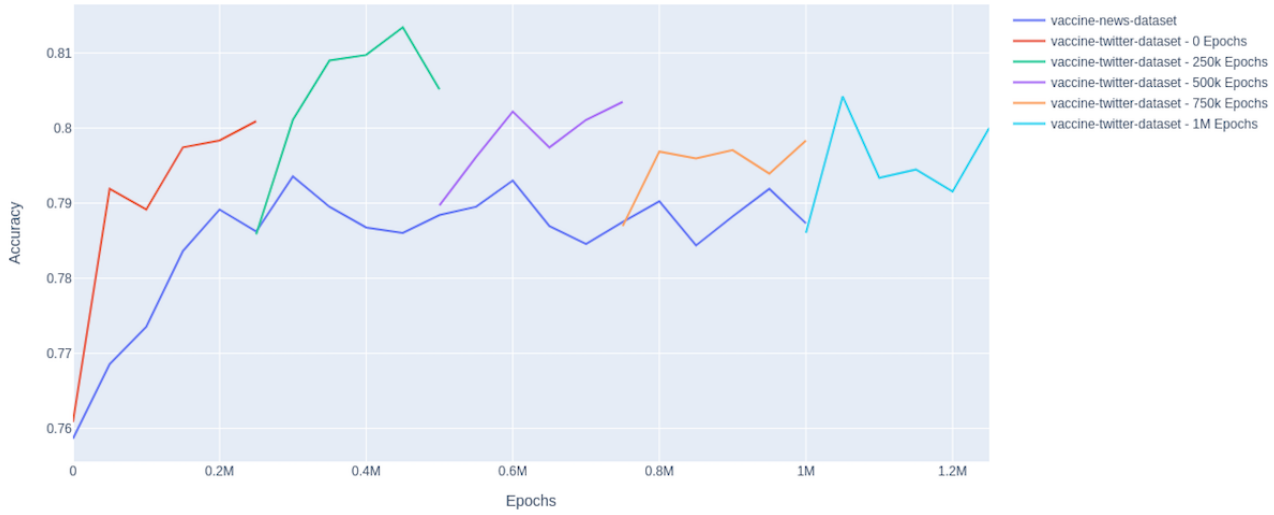


Table 2. Accuracy comparisons.

	Accuracy	*F* score
Coder average	0.795861	0.739396
Coder 1: EK	0.833211	0.796272
Coder 2: SCM	0.775165	0.710356
Coder 3: SD and CD	0.779206	0.711559
Neural network: no embeddings	0.436697	0.436697
Neural network: GloVe[a] word embeddings	0.544954	0.457813
LSTM[b]: no embeddings	0.631193	0.549997
LSTM+GloVe word embeddings	0.655046	0.593942
BERT[c]: default weights	0.766972[d]	0.718878
BERT: domain-specific	0.818349[d]	0.775830

[a]GloVe: Global Vectors for Word Representation.

[b]LSTM: long short-term memory.

[c]BERT: Bidirectional Encoder Representations from Transformers.

[d]The final accuracy scores for the Bidirectional Encoder Representations from Transformers–based models are based on selecting the best network from the results based on the results from the development data set. The reported numbers are from evaluating the training data set.

Discussion

Principal Findings

The categories chosen in this study underwent several revisions to ensure that they could be clearly understood. The annotators were fluent English speakers with a postgraduate degree and several years of work experience in the field.

Some of the nuances contained in the tweets meant that it was difficult to categorize them as definitively one stance. Thus, even when the same coder was asked to code a nearly identical tweet at a later time, they chose the same code four out of five times. After being given a second opportunity to code all duplicate tweets that had inconsistencies, the annotators met and discussed the categories that they disagreed on. The average final accuracy was then 79.6%.

Ideally, the correct annotating should be the annotating that an average of a large number of experienced annotators would have chosen. Limiting the number of annotators to 3 resulted in cases where all annotators by chance coded the same tweet identically and erroneously and cases where none of the annotators chose the categories that a larger number of annotators would have chosen. It is therefore reasonable to assume that the accuracy of 79.6% might be slightly optimistic in terms of what can be expected from an average human coder, even one who has long experience in the area.

The task of annotating is challenging because it is open for interpretation, which is a challenge that NLP also struggles with. Our tests showed that a simple neural network that had no problem achieving an accuracy of more than 80% on an IMDb movie review task was unable to predict anything better than the most prevalent category when it was tested on maternal vaccination tweets.

Unsurprisingly, LSTM networks perform better than ordinary neural networks. Pretrained embeddings help in all cases. Using GloVe word embeddings increases the accuracy. With the LSTM network, we achieved an accuracy of 63.1% and improved this to 65.5% by adding pretrained GloVe word embeddings. However, LSTM networks still lag behind what could be considered human accuracy.

In contrast, transformer-based architectures perform significantly better. By using the pretrained openly available BERT weights, we achieved an accuracy of 76.7%. This is approximately the same level of accuracy achieved by the coder with the lowest accuracy in this study.

Domain-specific pretraining also shows potential. Although pretraining does require some computing power, it does not require manual annotating, which could entail high costs in terms of time and resources. It is also worth noting that we deliberately trained the model only on general vaccine terms. We did not perform optimizations specifically for the domain of maternal vaccines. The main reason for this is that we wanted weights that were transferable to other tasks in the field of vaccines.

In our setting, the best result of 81.8% accuracy was achieved after initial training on news articles about vaccines and then training on vaccine-related tweets. This accuracy is better than the average of the 3 annotators, even after the annotators had carried out multiple annotatings of the same tweet and had been given the opportunity to recode any inconsistencies.

In our opinion, it is doubtful that any individual coder would achieve more than 90% accuracy on this task simply because it is difficult, even with a much larger number of annotators, to agree on an absolute categorization. There will always be tweets that are open to interpretation, preventing a hard target of absolute accuracy.

Limitations

We used a limited data set, especially for the tweet data set containing only 1 million vaccine-related tweets. It is also reasonable to assume that pretraining on a larger data set of non–vaccine-specific tweets could have a positive effect because the language of tweets is quite different from that of other texts. Enlarging the data sets is an easy way of potentially increasing the accuracy.

Although they are accurate, transformer-based models are demanding to run to analyze large amounts of text. This could be a challenge when used for monitoring purposes. However, most likely, this is a problem that will lessen in the future.

After Google released BERT late in 2018, there have been multiple general improvements made by Facebook, Microsoft, and Google to the model's transformer-based architecture to improve the base models [15,18,19]. These have not been implemented in this study. There is currently significant research activity in the field, and it is reasonable to assume that implementing these improvements in the base model and restarting the domain-specific pretraining checkpoints would lead to higher accuracy in our categorization.

Conclusions

Being able to categorize and understand the overall stance in social media conversations about vaccinations, especially in terms of identifying clusters of discouraging or ambiguous conversations, will make it easier to spot activities that may signal vaccine hesitancy or a decline in vaccine confidence with greater speed and more accuracy. To manually, and continually, monitor these conversations in today's information society is near impossible. In that respect, it has always been obvious that NLP has huge potential because it can process an enormous amount of textual information.

However, so far, NLP has only been able to solve very easy tasks and is unable to handle the nuances in language related to complicated issues (eg, attitudes toward vaccination). The new advances in transformer-based models indicate that they are about to become a useful tool in this area, opening up a new area for social research.

We have demonstrated that with a training data set of approximately 1600 tweets, we were able to obtain at least the accuracy that should be expected of a trained human coder in categorizing the stance of maternal vaccination discussions on social media. Although there are benefits to increasing this accuracy even more, the main research challenge is to reduce the number of training samples. So far, this has been an

underprioritized area of research and an area where we should expect advances in the future. The real benefit from the technology will first be apparent when we are able to do this kind of categorization with only a few initial examples. Being able to categorize text in large corpora gives us a new tool for tracking and ultimately understanding vaccine stance and sentiment.

Acknowledgments

PEK was funded by the European Commission for the call H2020-MSCA-IF-2017 and the funding scheme MSCA-IF-EF-ST for the Vaccine Media Analytics project (grant agreement ID: 797876).

HJL, PP, SM, SD, and EK were funded by a grant from GlaxoSmithKline to support research on maternal vaccination.

The funders had no role in the study design, data collection, analysis, interpretation, or writing of this paper. This research was supported with Cloud TPUs (Tensor Processing Units) from Google's TPU Research Cloud.

Conflicts of Interest

HJL's research group, the Vaccine Confidence Project (HJL, PP, SM, SD, and EK) received research funding from GlaxoSmithKline, Merck, and Johnson & Johnson. HJL served on the Merck Vaccines Strategic Advisory Board and received honoraria for participating in GSK training sessions. None of the other authors have any conflicts of interest to declare.

Multimedia Appendix 1
Maternal vaccination keyword search.
[DOCX File , 418 KB - medinform_v9i10e29584_app1.docx]

Multimedia Appendix 2
Vaccination keyword search terms.
[DOCX File , 7 KB - medinform_v9i10e29584_app2.docx]

References

1. Wilcox CR, Bottrell K, Paterson P, Schulz WS, Vandrevala T, Larson HJ, et al. Influenza and pertussis vaccination in pregnancy: portrayal in online media articles and perceptions of pregnant women and healthcare professionals. Vaccine 2018 Nov 29;36(50):7625-7631 [FREE Full text] [doi: 10.1016/j.vaccine.2018.10.092] [Medline: 30401620]
2. Kang GJ, Ewing-Nelson SR, Mackey L, Schlitt JT, Marathe A, Abbas KM, et al. Semantic network analysis of vaccine sentiment in online social media. Vaccine 2017 Jun 22;35(29):3621-3638 [FREE Full text] [doi: 10.1016/j.vaccine.2017.05.052] [Medline: 28554500]
3. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 2018 Oct;108(10):1378-1384. [doi: 10.2105/AJPH.2018.304567] [Medline: 30138075]
4. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011 Oct;7(10):e1002199 [FREE Full text] [doi: 10.1371/journal.pcbi.1002199] [Medline: 22022249]
5. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl Based Syst 2018 Dec;161:124-133. [doi: 10.1016/j.knosys.2018.07.041]
6. Smith MC, Dredze M, Quinn SC, Broniatowski DA. Monitoring real-time spatial public health discussions in the context of vaccine hesitancy. 2017. URL: http://ceur-ws.org/Vol-1996/paper2.pdf [accessed 2021-08-26]
7. Mohammad S, Sobhani P, Kiritchenko S. Stance and Sentiment in Tweets. ACM Trans Internet Technol 2017 Jul 14;17(3):1-23 [FREE Full text] [doi: 10.1145/3003433]
8. Meltwater homepage. Meltwater. URL: https://www.meltwater.com/ [accessed 2021-08-26]
9. VACMA GitHub repository. GitHub. URL: https://github.com/peregilk/VACMA-PUBLIC [accessed 2021-08-26]
10. Martin S, Kilich E, Dada S, Kummervold PE, Denny C, Paterson P, et al. "Vaccines for pregnant women…?! Absurd" - mapping maternal vaccination discourse and stance on social media over six months. Vaccine 2020 Sep 29;38(42):6627-6637. [doi: 10.1016/j.vaccine.2020.07.072] [Medline: 32788136]
11. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Preprint posted online October 11, 2018 [FREE Full text]
12. SQuAD2.0: The Stanford Question Answering Dataset. SQuAD. URL: https://rajpurkar.github.io/SQuAD-explorer/ [accessed 2021-08-26]
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. arXiv. Preprint posted online June 12, 2017 [FREE Full text]

14. Tensorflow code and pre-trained models for BERT. GitHub. 2020. URL: https://github.com/google-research/bert [accessed 2021-08-26]
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 July 26, 2019 [FREE Full text]
16. Chollet F. Deep Learning with Python, 1st Edition. Greenwich, CT, USA: Manning Publications Co; 2017.
17. Dhande L, Patnaik G. Analyzing sentiment of movie review data using Naive Bayes neural classifier. Int J Emerg Trends Technol Comput Sci 2014;3(4):1-8 [FREE Full text]
18. He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv. Preprint posted online June 5, 2020 [FREE Full text]
19. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. arXiv. Preprint posted online September 26, 2019 [FREE Full text]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
GloVe: Global Vectors for Word Representation
LSTM: long short-term memory
NLP: natural language processing
TPU: Tensor Processing Unit

Edited by C Lovis; submitted 13.04.21; peer-reviewed by X Cheng, A Fernandes; comments to author 02.07.21; revised version received 15.07.21; accepted 19.07.21; published 08.10.21.

Please cite as:
Kummervold PE, Martin S, Dada S, Kilich E, Denny C, Paterson P, Larson HJ
Categorizing Vaccine Confidence With a Transformer-Based Machine Learning Model: Analysis of Nuances of Vaccine Sentiment in Twitter Discourse
JMIR Med Inform 2021;9(10):e29584
URL: https://medinform.jmir.org/2021/10/e29584
doi:10.2196/29584
PMID:34623312

Original Paper

Building a Shared, Scalable, and Sustainable Source for the Problem-Oriented Medical Record: Developmental Study

Christophe Gaudet-Blavignac[1,2], BSc, MSc; Andrea Rudaz[3], MHSA, MD; Christian Lovis[1,2], MPH, MD

[1]Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland

[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[3]Medical and Quality Directorate, Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Christophe Gaudet-Blavignac, BSc, MSc
Division of Medical Information Sciences
Geneva University Hospitals
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 223726201
Email: christophe.gaudet-blavignac@hcuge.ch

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2022/8/e41257

Abstract

Background: Since the creation of the problem-oriented medical record, the building of problem lists has been the focus of many studies. To date, this issue is not well resolved, and building an appropriate contextualized problem list is still a challenge.

Objective: This paper aims to present the process of building a shared multipurpose common problem list at the Geneva University Hospitals. This list aims to bridge the gap between clinicians' language expressed in free text and secondary uses requiring structured information.

Methods: We focused on the needs of clinicians by building a list of uniquely identified expressions to support their daily activities. In the second stage, these expressions were connected to additional information to build a complex graph of information. A list of 45,946 expressions manually extracted from clinical documents was manually curated and encoded in multiple semantic dimensions, such as International Classification of Diseases, 10th revision; International Classification of Primary Care 2nd edition; Systematized Nomenclature of Medicine Clinical Terms; or dimensions dictated by specific usages, such as identifying expressions specific to a domain, a gender, or an intervention. The list was progressively deployed for clinicians with an iterative process of quality control, maintenance, and improvements, including the addition of new expressions or dimensions for specific needs. The problem management of the electronic health record allowed the measurement and correction of encoding based on real-world use.

Results: The list was deployed in production in January 2017 and was regularly updated and deployed in new divisions of the hospital. Over 4 years, 684,102 problems were created using the list. The proportion of free-text entries decreased progressively from 37.47% (8321/22,206) in December 2017 to 18.38% (4547/24,738) in December 2020. In the last version of the list, over 14 dimensions were mapped to expressions, among which 5 were international classifications and 8 were other classifications for specific uses. The list became a central axis in the electronic health record, being used for many different purposes linked to care, such as surgical planning or emergency wards, or in research, for various predictions using machine learning techniques.

Conclusions: This study breaks with common approaches primarily by focusing on real clinicians' language when expressing patients' problems and secondarily by mapping whatever is required, including controlled vocabularies to answer specific needs. This approach improves the quality of the expression of patients' problems while allowing the building of as many structured dimensions as needed to convey semantics according to specific contexts. The method is shown to be scalable, sustainable, and efficient at hiding the complexity of semantics or the burden of constraint-structured problem list entry for clinicians. Ongoing work is analyzing the impact of this approach on how clinicians express patients' problems.

(JMIR Med Inform 2021;9(10):e29174) doi:10.2196/29174

KEYWORDS

medical records; problem-oriented; electronic health records; semantics

Introduction

Background

The concept of a problem-oriented medical record is as old as 1968 [1]. One of the key elements of this approach is a list of relevant problems, current or past, which are important for understanding the patient's condition. A problem can be anything, complaints, symptoms, existing or previous conditions, diagnosis, procedures, socioeconomic issues, etc. This list is the corner stone of the clinician's education and the patient record. It is used from the first encounter, where it is named *chief complaint* to drive clinical reasoning but increasingly to support electronic decision support and diagnostic or care pathways. With the widespread adoption of electronic health records (EHRs) and since the Meaningful Use Act established the problem list as a requirement for care facilities [2,3], it has been the focus of much research and multiple improvements. However, its digitization has brought new opportunities and challenges. Problem lists vary in time and are influenced by the conditions of the population a care facility deserves or the specialties covered.

Creation and Evolution

The building of a problem list can be driven by free-text entries made by clinicians or by the creation of a finite list of items from which they can choose. Terms included in these premade lists are often taken from existing terminology. Compared with the use of free text, a premade list allows for more structured data and easier secondary use [4-6]. The use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [7,8] for the problem list seems to provide good coverage [9-11]. In 2009, the Clinical Observations Recording and Encoding Problem List Subset of SNOMED CT was created using data from 8 institutions [12]. The terms extracted from those institutions were mapped to the SNOMED CT concepts to create a subset usable as a problem list. The current version contains 6565 SNOMED CT concepts. Other approaches have been explored, such as the automatic generation of a patient's problem list using natural language processing and international terminologies but with lists of less than 200 problems focused on a clinical specialty [13-17]. When creating a problem list, the equilibrium between a list representing what care professionals need to express and an interoperable controlled vocabulary is difficult to find [18].

Using terminology such as the International Classification of Diseases, 10th revision (ICD-10) [19] as a source of expression for a problem list can lead to multiple issues. A classification is a partition of reality in a finite set of categories, resulting in a phenomenon called residual aggregation or residual category. For example, *Other specified immunodeficiencies*; *Disorder of pancreatic internal secretion, unspecified*; or even *Fracture of unspecified phalanx of other finger* exist in ICD-10 to cover all concepts that do not fall in another category [20]. This type of term is not suited for a problem list because it does not represent problems that clinicians can reasonably enter.

Another challenge in using a classification as a source is based on its organization, tightly connected to the *intention*, which supported its development. For example, ICD-10 aims to properly express causes of deaths and morbidity, the International Classification of Primary Care, 2nd edition (ICPC-2) [21] focuses on primary care problems, and the Logical Observation Identifiers Names & Codes [22] covers observations and laboratories. Thus, each of these have a specific structure and a dedicated organization of their hierarchy to answer the requirements of their use.

The problem list should be able to represent any of those *intentions*, regardless of their future interpretations according to specific classificatory intentions, and without restricting elements to only one classification nor requiring clinicians to know the organization of all of them. A hierarchy such as the ICD-10 results in choices that will favor some dimensions over others. As an example, there is no infectious disease chapter in ICD-10, which seriously complicates the identification of infectious diseases. As a consequence, our approach focuses on using real-world clinicians' expressions as the primary source and then manually adding as many *semantically meaningful* dimensions as needed.

Maintenance and updating of problem lists is also challenging. For example, during the current pandemic, it was suddenly required to add several new entries to express the specific spectrum of COVID-19. Such rapid adaptations of the list must be rapidly implemented and should not depend on the update cycle of an international classification.

Using an efficient problem list requires considerable background information. For example, the same problem can be addressed several times. This is sometimes appropriate, such as repeated fractures, and sometimes inappropriate, such as repeating at each encounter that the patient has hypertension. Describing the semantics properly facilitates and speeds up the work of clinicians [23,24]. Semantic dimensions should support recognition and reconciliation algorithms and different views of the list, by specialty, organ, and severity, to name a few [25,26] or to support graph-based, symbolic, machine learning, or clustering algorithms to group concepts along a navigation that answers the needs of clinicians, case managers, researchers, etc [27].

Implementation and Adoption

Although the advantages of a well-maintained problem list are clear, numerous issues have been raised in the way it should be implemented. Engaging users to document a list of problems for their patients in a complete and efficient manner is a challenge. Clinicians in hospitals are under constant pressure, and the effort to pivot from a free-text problem list to a dedicated EHR module can be important. Factors such as gap reporting,

problem-oriented charting, or links to billing codes have shown some positive impact on the completeness of the list documented by clinicians [23,28]. In addition, training and education seem to be key factors in adoption [29-31]. In 2016, Simons et al [32] proposed a list of determinants for the successful implementation of a problem-oriented medical record. It includes the completeness, interoperability, usability, and training of staff.

In this paper, we aim to address the challenges of building and implementing a shared, multipurpose common problem list at Geneva University Hospitals (HUG) using an approach based on the clinician's language and semantic dimension encoding. The driving concepts of this work are that the content of the list should be created with the care professionals to match their needs and that the list should be mapped to terminologies to (1) improve adoption, with metadata for completers and (2) for secondary use of data [4,33]. After the description of the building, implementation, and iterative improvements of the list, an analysis of its use over 4 years is presented.

Methods

Approach

This approach focused on 2 goals. First, it allows clinicians to express themselves freely with a list representing the language used every day in clinical interactions and working with a free-text completer rather than a constrained closed list. Second, it allows the use of the list for multiple purposes in the hospital, other than supporting the care activities of the clinicians. The latter is performed by a back-office multidimensional extension of metadata of free-text expressions.

Common List Creation

To represent the language of the clinicians, the starting point is sentences expressing problems written by clinicians. The initial list was created based on 2 sets of documents extracted from the HUG's data warehouse, one from the internal medicine department and the other from the surgery department. Each set was composed of 10,000 admission letters and 10,000 discharge summaries for a total of 40,000 documents. Every natural language sentence in these documents was extracted using automatic tools without further processing. Those sentences were then manually selected if they represented a potential candidate, curated for typos and grammatical normalization such as plural or uppercase reserved for proper names. The abbreviations have been expanded but kept. Rules applied to build this list were inclusive, covering problems of any type, including but not limited to medical, surgical, socioeconomic, psychologic, logistic, etc. Synonymy is allowed, so that multiple expressions expressing the same problem are present, such as *generalized pain* and *pain everywhere* but connected as synonyms. Every granularity is allowed as long as the expression is used by clinicians. The only strict rule is that an expression must be syntactically and morphologically unique.

The list of expressions is improved based on 2 axes: vertical (expressions) and horizontal (dimensions). Extensions of the list require deployment in a specific clinical context, for example, neurosurgery. In this case, discussion with clinicians and analysis of their clinical documents allows us to build a set of specific expressions for that context, which are added to the common list before the deployment. Adjustments of the list are also iteratively made based on use, aided by the fact that the problem list management module is based on a syntactic completer allowing clinicians to enter free text and then select an expression if appropriate, or keep the original free text. The modifications of the list, expressions, and activity state of expressions are fully historicized based on use. Deletions are usually forbidden, which happened only once after a one-year evaluation of the impact of deleting entries: ensuring they had never been used and the impact of their absence on tools such as completers, parsers and colocations, word embedding, etc.

A monthly use analysis with all expressions chosen, by whom, in which context and the potential free text added are used to improve the list.

Semantic Dimensions

Overview

We considered a semantic dimension as any metadata added to the list of expressions to improve its use for a specific purpose. This purpose can be the completer functionalities, for example, for ambiguous abbreviations (in French, *TV* can mean *tachycardie ventriculaire* or *toucher vaginal*), or when the expression is gender specific, such as all expressions relating to *prostata*. Some dimensions are related to national classifications, such as the Swiss Classification for Surgical Interventions (CHOP) [34], or international classifications, such as ICPC-2 or ICD-10, including their various versions (several releases of ICD-10, for example). Finally, some are internal to the organization, such as a specific identification for surgery requiring a surgical theater, used for logistics and resource management at HUG. Expressions can have no or several entries in any specific dimension.

Encoding was performed by domain experts. For example, the ICD-10 and CHOP classifications have been made by a coding expert of the billing division of HUG, SNOMED CT encoding by a physician, ICPC-2 encoding by an outpatient physician, etc. Several dimensions, such as chronic or acute, gender specificity, and syntactical dimensions, have been conducted by medical students.

The dimensions described here are not exhaustive but representative. The coding of the dimensions is a complex activity, mostly toward maintaining global coherence. In this work, the strategy is to have a clear definition of a dimension and aim to reach the best quality of representation of that dimension, regardless of the others except the expression itself. The objective is that a specific expression that can be represented in that dimension must be represented with the highest precision possible for that dimension, respecting only the rules specific to it. This strategy has several advantages. It allows to keep the intention of the dimension to be coded at best and allows the encoding work to be distributed among several actors, domain experts, or students, according to their competences specific to that dimension and their understanding of the expression. Finally, a specific expression can be understood differently and with a different granularity, according

to the perspective of the dimension used, or seen as the sum of some or all dimensions.

General Classifications

International Classifications

ICD-10 is the basis for the billing of inpatient stays in Switzerland. Once every 2 years, the Swiss Confederation publishes its own version of the ICD-10 classification, which is a translation of the ICD-10, German Modification (ICD-10 GM), which is a slightly modified version of the ICD-10 released by the World Health Organization (WHO) [35]. Every expression in the list was first encoded with the ICD-10 WHO version to evaluate gaps in the list and perform subset definitions for specific use cases. Second, the list was encoded using the Swiss ICD-10 GM version. The encoding was performed using the official coding rulebook for hospital stays in Switzerland [36]. This dimension has been added in the aim of performing automatic coding of inpatient stays for billing, prediction tools for problems versus diagnosis, or support of pathways. Several versions of the ICD-10 are encoded, according to years, or to the source, including the WHO's original ICD-10.

ICPC-2 is a classification used to encode general practice clinical activities and primary care. It belongs to the WHO family of international classification [21]. This classification was chosen by the clinicians from the outpatient clinics for its ability to classify problems in simple categories relevant for care, such as symptoms, diagnosis, screening, or procedures. In addition to the activities of outpatient clinics, including research, this classification is used to generate alerts when adding multiple problems with the same ICPC-2 encoding.

The *Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)* is a term with more than 340,000 concepts and 1 million relationships [7,37]. It is described as the most comprehensive clinical health care terminology in the world and has become central to semantic interoperability. It has been chosen as the United States standard for encoding diagnoses and problem lists [38]. SNOMED CT includes powerful features such as the combination of concepts (postcoordination) or the expression constraint language, which can be used to perform complex queries on SNOMED CT encoded data. SNOMED CT is one of the pillars for the semantically driven activities for data science at HUG and allows the connection of many different aspects of the EHR, such as problem lists, formularies, and other structured data. It allows complex queries, such as every problem related to an organ, or including an inflammatory process. Owing to the size and complexity of the terminology, encoding several expressions in SNOMED CT requires a significant amount of time and experience. The encoding of the expressions uses only single or multiple precoordinated elements, a step toward fully postcoordinated expressions.

National Classifications

CHOP [34] is used to encode and bill surgical interventions. Goals similar to those of the ICD-10 GM were added. Every expression in the list that can be mapped to a CHOP code was mapped and updated annually when the new version was released.

Internal Classifications

Several internal classifications are used in specific contexts, which are illustrated hereafter.

Department or Specialty-Specific Lists

Adult and pediatric emergency departments use specific problem lists, which were included in the process. Most of the time, these lists were derived from the ICD-10. Appropriate dimensions were added, including specialty preferences. The adaptations were systematically validated by specialty experts. The same process has been applied in several specialties, such as oncology and neurosurgery.

Clinical Decision Support

Some expressions and dimensions have been added specifically to support computerized provider order entry, exemplified with *antibiotic prescription support* to improve choice of antibiotics, monitor, and lower antibiotic resistance. The expressions related to that list were added, properly encoded, and their belonging to problems related to antibiotic prescription added in a new dimension, so that it could be used in several modules of the EHR.

Surgical Intervention List

One key development enabled the use of the list as a unique source of expressions for surgical intervention planning and documentation. When an intervention is planned in the hospital, it triggers a chain of events that will lead to the intervention. The operating room must be booked, staff must be appointed, specific devices and materials must be ordered, etc. Historically, this process was separated into silos, medical, paramedical, or logistic with separate lists. The list of surgical interventions used for operating room planning was manually integrated into the common list as a new dimension. This integration was made by specialties and is still ongoing. It allowed the common list to become a single source of expressions for surgical intervention planning.

Nutrition and Dietetic Diagnoses List

The most recent development of the list focused on the diagnoses used by the dieticians and nutritionists, which was a list of expressions extracted from the Terminologie Internationale de Diététique et de Nutrition [39]. These expressions were curated and integrated as a new dimension, making the common list the single source of expressions for the nutritionist and dieticians of the hospital.

Other Dimensions

Other specific dimensions are useful for numerous purposes. The gender specificity dimension defines whether an expression is gender specific, such as *vasectomy*. The intervention dimension defines an act performed and differentiates it from interventions requiring surgery theater. Multiple other dimensions are used for numerous purposes, such as possible abbreviations of the expression, preferred terms, chronic or acute, etc.

Language

The expressions being in French, an English translation was prepared, and keywords of the expressions were added in both French and English.

Results

Evolution of the List

The list of problems presented in this work had to *compete* with 17 specific, specialty vertical, local problem lists and was proposed as an additional choice for clinicians. They could freely choose between their *usual* lists and the new one. This competitive approach was a strong incentive to stick to the needs of clinicians and become their *preferred* list. Within the first year, the new list became the most used in most cases, and the legacy lists were then removed. The 2 first years required frequent adjustments, but with a slowing down frequency up to the current situation, which is on specific demand, such as COVID-19, or monthly. Table 1 summarizes the major releases.

Table 1. Major releases, corpus size, and comments.

Date of release	Active problems, n	Modifications
January 2017	45,946	• First production deployment
September 2017	45,458	• Partial integration of expression for surgery planning • Corrections of expressions
January 2018	51,255	• 5867 expressions created from legacy list use and free-text entries
February 2018	50,822	• Integration of expressions for antibiotics prescription and monitoring project • Corrections of expressions
May 2018	52,040	• 1091 expressions created from legacy list use and free-text entries
November 2018	52,211	• Integration of the list for adult emergency ward • Abbreviations system integration
August 2019	51,824	• 310 expressions created on demand from users
January 2020	52,956	• Integration of expressions for surgery planning • Integration of a list of diagnoses used by dieticians and nutritionists • Integration of the list for pediatric emergency ward
April 2020	52,958	• Emergency adding of 2 expressions for SARS-CoV-2 cases
August 2020	20,120	• Inactivation of 32,840 never used expressions • Preferred term system integration

In January 2017, the list was deployed in the geriatric and general pediatric division of the HUG, as well as part of the rehabilitation medicine division and ambulatory primary care division. The list was then progressively deployed in new divisions. Table 2 summarizes these deployments and Table 3 exposes some descriptive statistics of the current list.

Table 2. Deployment of the list in new divisions by date.

Date	Division
April 2017	Neurosurgery
May 2017	Neurology
May 2017	Visceral surgery
November 2017	Psychiatry (adult and pediatric)
November 2018	Rehabilitation
September 2019	Adult emergency
September 2020	Internal medicine
September 2020	Oncology
September 2020	Cardiology

Table 3. Some descriptive statistics of the list.

Type of expression or encoding	Expressions (N=20,120), n (%)
Active expressions	20,120 (100)
Abbreviations	2127 (10.57)
ICPC-2[a] encoding	20,120 (100)
ICD-10[b] WHO[c] 2008 encoding	11,860 (58.95)
ICD-10 GM[d] 2018 encoding	18,481 (91.85)
CHOP[e] 2019 encoding	1223 (6.08)
SNOMED CT[f] encoding	9222 (45.83)
Gender specificity encoding	805 (4)
Acute or chronic specificity encoding	8013 (39.83)
Intervention encoding	1855 (9.22)
Surgery planning	985 (4.89)
Antibiotic decision support	553 (2.75)
Adult emergency ward	1108 (5.51)
Pediatric emergency ward	939 (4.67)
Nutrition and dietetics	139 (0.69)

[a]ICPC-2: International Classification of Primary Care, 2nd edition.

[b]ICD-10: International Classification of Diseases, 10th revision.

[c]WHO: World Health Organization.

[d]ICD-10 GM: International Classification of Diseases, 10th revision, German Modification.

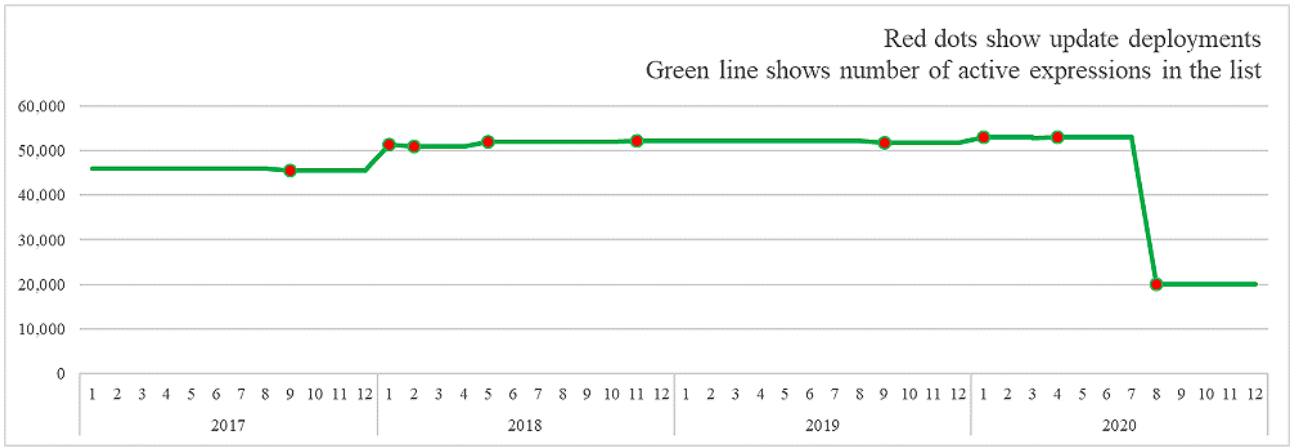
[e]CHOP: Swiss Classification for Surgical Interventions.

[f]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

For us, an important success indicator is that currently, 3 major divisions, internal medicine, geriatrics, and rehabilitation, decided to remove free-text entry possibility, judging that the common list was sufficiently complete for their use.

In 4 years, 7270 expressions were added from legacy lists, free-text, or users' requests. After 3 years of use, all 32,840 expressions that were never used or linked to any specific project were inactivated from the source and deleted for production. The current version of the list contains 20,120 active expressions. The evolution of the number of expressions in the list is shown in Figure 1.

Figure 1. Evolution of the number of active expressions in the common list.



Use of the List

After 4 years of use, all problems created were extracted from HUG's data warehouse, representing 1,146,135 problem creations. Among them, 59.69% (684,102/1,146,135) were chosen from the common list, 14.83% (169,970/1,146,135) from legacy lists, and 25.48% (292,063/1,146,135) entered as free-text entries. Over the legacy list problems, 63.01% (107,095/169,970) were created during the first year. In December 2017, the month with the largest proportion of free-text entries, 37.47% (8321/22,206) of the problems were created using this method. In December 2020, the last month of the observation period, 18.38% (4547/24,738) of the problems were created using free text and 80.18% (19,836/24,738) using the list.

From the common list, 15,232 distinct expressions were used at least once. Figure 2 shows the absolute number of problems created by the month and their origin. Legacy lists combine all problems arising from the 17 legacy lists in production in the HUG at the time of the first deployment and are progressively abandoned.

Figure 2. Number of problems created by month according to their origin.

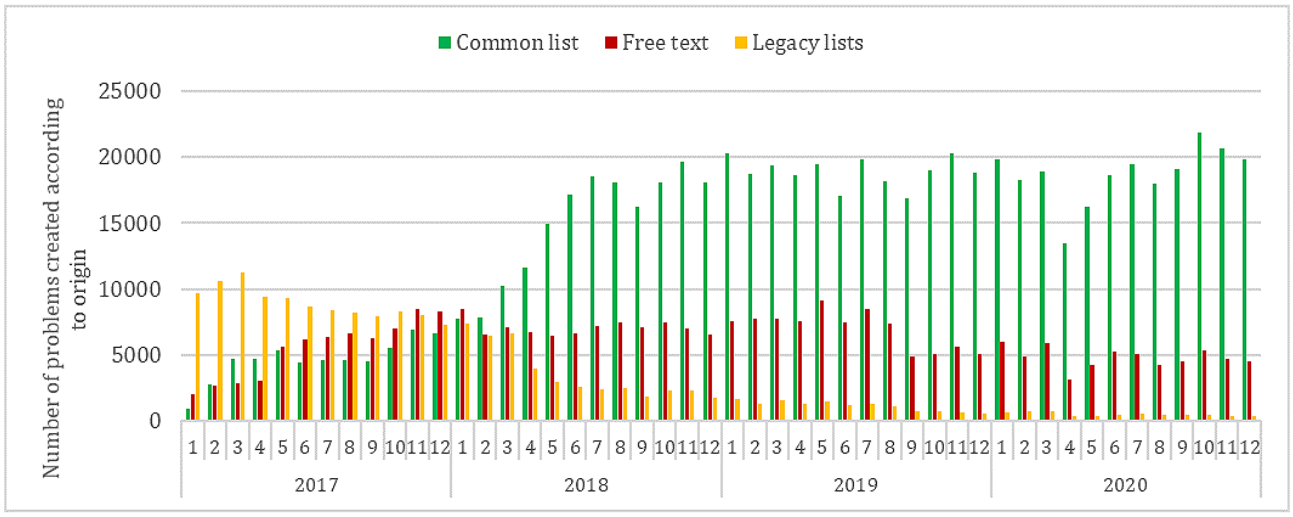
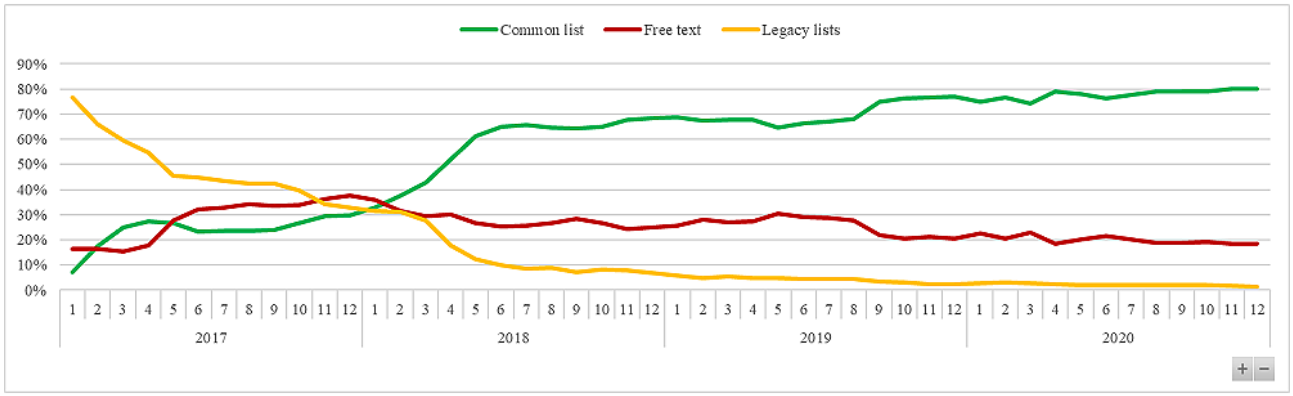


Figure 3 displays the proportion of problems chosen in the common list versus legacy lists and free-text entries.

Figure 3. Proportion of problems created by month according to their origin.



The 20 most frequently used expressions in the list are listed in Table 4. Free-text entries and expressions from the legacy lists were not included.

Table 4. The 20 most frequently used expressions of the common list over 4 years.

Expression	English translation	Uses, n
Hypertension artérielle	Arterial hypertension	16,974
Insuffisance rénale aiguë	Acute renal failure	6391
Hypercholestérolémie	Hypercholesterolemia	5219
Accouchement normal d'un nouveau-né vivant par voie basse	Normal vaginal delivery of a liveborn	5045
Appendicectomie	Appendicectomy	4550
Hypertension artérielle traitée	Treated arterial hypertension	4230
Décompensation cardiaque	Cardiac decompensation	4159
Douleur thoracique	Thoracic pain	3707
Hypokaliémie	Hypokalemia	3363
Troubles cognitifs	Cognitive disorder	3323
Hyponatrémie	Hyponatremia	3212
Infection à SARS-CoV-2 (COVID19)	SARS-CoV-2 (COVID-19) infection	3118
Fibrillation auriculaire	Atrial fibrillation	3055
Diabète type 2	Type 2 diabetes	3051
Insuffisance rénale chronique	Chronic renal failure	3002
Malnutrition protéino-énergétique grave	Serious protein-energy malnutrition	2898
Dyslipidémie	Dyslipidemia	2844
Obésité	Obesity	2833
Asthme	Asthma	2756
Douleur abdominale	Abdominal pain	2749

Finally, the list was exploited for various research activities, training machine learning models using various mappings, predicting billing codes of a stay using the ICD-10 encoding, or for workload predictions during the multiple waves of the pandemic. This work has not been discussed in this paper.

Discussion

Principal Findings

After 4 years of deployment and iterative improvements, a list of 20,120 active expressions mapped to more than 14 semantic dimensions was deployed in most major divisions of HUG and used to create 684,102 new problems. Specific dimensions allowed the list to be used for various purposes, such as surgical planification, decision support or nutrition, and dietetic diagnosis.

Manually building a problem list is a time-consuming task and starting from the clinician language as a source of expressions is a double-edged sword. It aims to improve information precision to support clinicians in finding the most appropriate expressions that best represent the conditions of patients. Cost tends to increase noise when proposing numerous expressions with small variation, syntactically, semantically, or both, depending on the completer used. This effect can be mitigated using the features of the dimensions, either syntactically, such as abbreviations and common variants, or semantically, by using the aggregation properties of classificatory dimensions. This allows us to search for the entirety of the list while reducing the number of possibilities proposed to the most pertinent set. These tools were not discussed in this work. Finally, statistics on the use of the list are important to improve it, such as progressively filtering out never used expressions or improving granularity in existing ones that are extended by free texts, for example.

We noticed that problems appearing frequently in practice tend to have multiple variations, with various levels of granularity or additional information, naturally improving the expressiveness of the list and the ease for clinicians to find the most appropriate element. On the other hand, rare problems tend to have fewer representations, if any in the list thus reinforcing the need to keep free-text entries.

The many dimensions that have been encoded allow the comparison of the list of expressions and their coverage for the respective coverage of classifications. For example, taking ICPC-2 and ICD-10, the immediate observation is that the list contains elements that can be expressed in both classifications, but in many more lexical variants. On the other hand, many classifications are not found in the list, for many of them not codable elements or unmet conditions in our setting. As a result, the list covers more than any of the classifications separately but only meaningful expressions. Moreover, it frees care professionals from the task of knowing multiple classifications and their structures. This reduces the compression of information while maintaining strong interoperable capabilities through semantic dimensions.

Semantic dimensions are a major addition of this approach. They bridge the need for various representations of a concept as expressed by clinicians with the need for semantic interoperability. By encoding each expression into all relevant dimensions, it was possible to reuse the created problems for other goals, for example, by extracting subsets related to a specific disease through ICD-10 encoding, all patients that undergo a specific procedure using the CHOP encoding or more complex queries such as all problems that include an inflammation process through the SNOMED CT encoding. However, the maintenance costs of these dimensions are important. The more dimensions there are, the more work it requires to add a new expression, as it must be encoded in possibly all of them. Moreover, classification updates (such as a new version of the ICD-10) sometimes require a full reading and update of the encoding.

The semantic dimensions linked to intrahospital use cases allowed the list to be used for multiple projects. Specific subsets for divisions, such as emergency wards, were beneficial for convincing users to start using the common list. The surgical planning addition promoted the list as a central source of expressions and concepts outside of the care domain. The role of the list as a central source of expressions for patients' problems is shown by the number of projects that included the addition of a dimension to the list. In a virtuous circle, the more the list was known, the more demands were made to adapt it to new needs.

As every project of this type, the final challenge is to convince users to use the module and teach them how to do so correctly. This has been heavily pushed in this work by the Medical and Quality Directorate, the team designing the problem list module in the HUG. Teaching both in person and through videos helped disseminate the use of the module in divisions that historically did not use it.

During the first year of deployment, the module was introduced and promoted in 4 new divisions of the hospital. This increases the number of users and the number of problems created. Those new users with no experience of the problem module are arguably the reason for the initial augmentation in the proportion of free-text entries seen in Figure 3. The diminution in problems created from the legacy list is to account for the progressive removal of those lists from the module. After this initial period, the proportion of free text diminishes progressively from 37.47% (8321/22,206) in December 2017 to 18.38% (4547/24,738) in December 2020, the lowest percentage in the full period. It is interesting to note that this period of 1 year also corresponds to the time it took for the common list to become the most used method for creating problems.

This reduction in the proportion of free-text entries shows that the common list corresponds to the needs of care professionals and that its adoption is progressing. Although it is not possible to determine the proportion of this evolution because of the content of the list, the functionalities of the problem module, or the dissemination effort, it seems likely that it is a combination of the three, and that only a transversal approach could succeed in this transition.

The situation before the deployment of the common list seemed preferable because the proportion of free-text entries was low and the use of legacy list was well-established. However, the final situation is arguably better for several reasons. First, the legacy lists lacked proper semantic interoperability. They were manually modified versions of existing classifications, with the limitations described before and the added complexity of manual, unverified modifications. They were not harmonized, and it was not possible to group or analyze problems from multiple lists without manual reading of the expressions. This prevented those lists from being used for other purposes, as the common list allows.

The apparent decrease in the number of problems created in April and May 2020 is explained by the COVID-19 pandemic. Indeed, the HUG stopped their elective activity and shifted to treating only patients with COVID-19, which reduced the number of patients with various problems and reduced the overall number of problems created.

Sustainability

Sustainability is an important aspect of large-scale projects, such as the creation of a common multipurpose problem list. For this specific issue, a common list presents interesting properties. As explained before, it can be extended vertically in 2 axes by adding new expressions and horizontally by adding new dimensions. This allows the list to quickly integrate new expressions, such as during the COVID-19 pandemic, or new dimensions such as dietetics and nutrition diagnoses. However, the amount of work required for vertical or horizontal extensions is not the same. A new expression can be encoded in all dimensions in a matter of minutes; however, in the worst case, the addition of a dimension requires going through every expression. Although, as the list has been kept to a manageable size by focusing only on expressions used in practice, this work can be performed with reasonable resources. Therefore, this list presents good flexibility and sustainability.

Reproducibility

The approach taken in this study was focused on the language of the clinicians. Therefore, the list of expressions is highly dependent on the clinicians, their language, their cultural background, and the population they cover. Therefore, the list itself will always be the most useful in the hospital where it has been created. However, the approach proposed to create the list is reproducible in any hospital wanting to create a problem list and for other use cases where a controlled vocabulary can be used but does not fit the language used in practice by caregivers.

Lessons Learned

This work allowed us to draw significant learning for the building and implementation of a problem list. These are listed in Textbox 1.

Textbox 1. Key learnings.

Key learnings

- Existing controlled vocabularies are too narrow or subject oriented to be used natively as problem lists.
- It is possible to build a problem list starting from the clinician's language to better match their needs.
- It is possible to reduce the expressivity needed for a problem list to a meaningful set of expressions used in practice.
- On purpose semantic dimension encoding allows secondary use of data.
- Internally building a list of expressions allows flexibility and quick adjustments when needed.

Limitations

Although the data and analysis included in this work were carefully carried out, some limitations are worth noting. First, the evaluation data were analyzed as a source of problems created. However, this does not translate to the complexity of the deployment of the list in the hospital. Indeed, the problem module is deployed in the EHR globally, but some divisions use it, while others do not. Inside these divisions, some teams of residents are more used to the module than others. Additional data should be gathered to track the dissemination effort, the training provided, to understand when the module was adopted in which division and by whom.

Finally, the proportion of common list, free text, and legacy list problems is only a proxy for user preferences. It does not account for other elements, such as division-specific guidelines or orally transmitted habits. To credit the progression of the common list to its quality is a conclusion that should be confirmed by a closer evaluation, in partnership with the users.

Conclusions

Overall, there is still room for improvement when building and implementing a problem list in the production environment of care. Most of the existing efforts use terms from existing terminology rather than focusing on the language used by clinicians. The perfect problem list that contains what care professionals want and can be used for every other use-case is yet to be created.

The proposed approach breaks with common approaches for the building of problem lists by directly addressing the gap between existing controlled vocabularies and real clinicians' language when expressing a patient's problem. Second, it brings new perspectives for secondary use by encoding the expressions in various semantic dimensions, allowing specific uses of the list in the hospital and beyond.

By applying this approach, more than 50,000 expressions were manually curated into a common problem list integrated in the EHR. Through iterative updates, the list was enriched and refined to 20,120 active expressions matching users' needs. More than 14 semantic dimensions were added to the list, including 5 major classifications and multiple dimensions internal to the hospital, such as division-specific adaptations, surgical planning, antibiotic prescription support, nutrition, and dietetic diagnoses. These additions pushed the adoption of the common list as a central, harmonized source of expression in the hospital. The recent decision of 3 major divisions of the hospital to remove the option to make free-text entries shows that the list corresponds to the needs of the users.

Manually creating and updating a set of expressions directly extracted from clinical documents has succeeded in HUG to engage users in transitioning from legacy systems to a new module including the common list. The overall number of problems created is increasing, while the problems entered as free text are decreasing.

The manual work required to build and maintain the list is substantial in the 3 domains, maintenance of the expressions, development of the problem module, and dissemination of its use. However, this approach provides a solution for keeping data interoperable while not constraining the user and allowing multiple use cases.

Moreover, with the large adoption of the list in the HUG, new perspectives open and new types of projects are possible. Ongoing developments include oncologic diagnoses with the addition of a dimension mapping expression to the third edition of the International Classification of Disease for Oncology, extension of the surgical planning dimension, or creation of the 2021 version of the ICD-10 GM and CHOP dimensions. The addition of dimensions for new international classifications, such as the eleventh revision of the ICD, are also evaluated. However, the improvement of the SNOMED CT dimension is currently prioritized over these additions, owing to the quantity of information expressible in SNOMED CT, the multiple mappings existing between SNOMED CT and other controlled vocabularies, and the national recommendation of this terminology for interoperability in Switzerland.

In addition, the common list allows new research projects in the medical domain, such as analysis of the problems documented for patients with COVID-19 or focusing on the language, such as the study of the search terms entered by clinicians compared with the problem selected in the list.

An evaluation of the impact of the list on the workload of clinicians and on the secondary uses of the produced data should be made to further validate the approach.

Acknowledgments

This project was partly funded by the Evolving Language National Centre of Competence in Research of the Swiss National Fund, N-603-11-01. The common problem list is available under the Creative Commons CC BY-SA 4.0 license on Yareta, the digital solution of the University of Geneva for archiving and preserving research data [40].

Conflicts of Interest

None declared.

References

1. Weed LL. Medical records that guide and teach. N Engl J Med 1968 Mar 14;278(11):593-600. [doi: 10.1056/NEJM196803142781105] [Medline: 5637758]
2. Office of the National Coordinator for Health Information Technology. Index for Excerpts from the American Recovery and Reinvestment Act of 2009 (ARRA). 2009. URL: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf [accessed 2021-01-04]
3. Holmes C. The problem list beyond meaningful use. Part I: The problems with problem lists. J AHIMA 2011 Feb;82(2):30-34. [Medline: 21337850]
4. Acker B, Bronnert J, Brown T, Clark JS, Dunagan B, Elmer T, et al. Problem list guidance in the EHR. J AHIMA 2011 Sep;82(9):52-58. [Medline: 21980907]
5. Elkin PL, Mohr DN, Tuttle MS, Cole WG, Atkin GE, Keck K, et al. Standardized problem list generation, utilizing the Mayo canonical vocabulary embedded within the Unified Medical Language System. Proc AMIA Annu Fall Symp 1997:500-504 [FREE Full text] [Medline: 9357676]
6. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. Proc AMIA Symp 1998:795-799 [FREE Full text] [Medline: 9929328]
7. SNOMED International SNOMED CT Browser. SNOMED International. URL: https://browser.ihtsdotools.org/? [accessed 2020-01-06]
8. SNOMED - Home. SNOMED International. URL: https://www.snomed.org/ [accessed 2021-01-07]
9. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. AMIA Annu Symp Proc 2003:699-703 [FREE Full text] [Medline: 14728263]
10. Penz JF, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, et al. Evaluation of SNOMED coverage of Veterans Health Administration terms. Stud Health Technol Inform 2004;107(Pt 1):540-544. [Medline: 15360871]
11. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006 Jun;81(6):741-748. [doi: 10.4065/81.6.741] [Medline: 16770974]
12. The CORE Problem List Subset of SNOMED CT®. U.S. National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed 2021-02-25]
13. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak 2005 Aug 31;5:30 [FREE Full text] [doi: 10.1186/1472-6947-5-30] [Medline: 16135244]
14. Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, et al. Building an automated problem list based on natural language processing: lessons learned in the early phase of development. AMIA Annu Symp Proc 2008 Nov 06:687-691 [FREE Full text] [Medline: 18999050]
15. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. Int J Med Inform 2008 Sep;77(9):602-612. [doi: 10.1016/j.ijmedinf.2007.12.001] [Medline: 18280787]
16. Meystre S, Haug P. Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. AMIA Annu Symp Proc 2006:554-558 [FREE Full text] [Medline: 17238402]
17. Meystre SM, Haug PJ. Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annu Symp Proc 2005:525-529 [FREE Full text] [Medline: 16779095]
18. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. Proc AMIA Symp 1998:280-284 [FREE Full text] [Medline: 9929226]
19. ICD-10 Version:2019. World Health Organization. 2019. URL: https://icd.who.int/browse10/2019/en [accessed 2021-01-07]
20. Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. J Biomed Inform 2019 Jun;100S:100002 [FREE Full text] [doi: 10.1016/j.yjbinx.2019.100002] [Medline: 34384571]
21. International Classification of Primary Care, Second edition (ICPC-2). World Health Organization. URL: https://www.who.int/standards/classifications/other-classifications/international-classification-of-primary-care [accessed 2020-11-25]
22. LOINC from Regenstrief. URL: https://loinc.org/ [accessed 2021-01-07]
23. Wright A, McCoy AB, Hickman TT, Hilaire DS, Borbolla D, Bowes WA, et al. Problem list completeness in electronic health records: a multi-site study and assessment of success factors. Int J Med Inform 2015 Oct;84(10):784-790 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.06.011] [Medline: 26228650]

24. Wang EC, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. J Am Med Inform Assoc 2020 Aug 01;27(8):1190-1197 [FREE Full text] [doi: 10.1093/jamia/ocaa125] [Medline: 32620950]
25. Hier DB, Pearson J. Two algorithms for the reorganisation of the problem list by organ system. BMJ Health Care Inform 2019 Dec;26(1):100024 [FREE Full text] [doi: 10.1136/bmjhci-2019-100024] [Medline: 31848142]
26. Hammond KW, Helbig ST, Benson CC, Brathwaite-Sketoe BM. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. AMIA Annu Symp Proc 2003:269-273 [FREE Full text] [Medline: 14728176]
27. Kreuzthaler M, Pfeifer B, Ramos JA, Kramer D, Grogger V, Bredenfeldt S, et al. EHR problem list clustering for improved topic-space navigation. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):72 [FREE Full text] [doi: 10.1186/s12911-019-0789-9] [Medline: 30943968]
28. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, McLoughlin KS, et al. Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial. J Am Med Inform Assoc 2012;19(4):555-561 [FREE Full text] [doi: 10.1136/amiajnl-2011-000521] [Medline: 22215056]
29. Bredfeldt CE, Awad EB, Joseph K, Snyder MH. Training providers: beyond the basics of electronic health records. BMC Health Serv Res 2013 Dec 02;13:503 [FREE Full text] [doi: 10.1186/1472-6963-13-503] [Medline: 24295150]
30. Bakel LA, Wilson K, Tyler A, Tham E, Reese J, Bothner J, et al. A quality improvement study to improve inpatient problem list use. Hosp Pediatr 2014 Jul;4(4):205-210. [doi: 10.1542/hpeds.2013-0060] [Medline: 24986988]
31. Klappe ES, de Keizer NF, Cornet R. Factors influencing problem list use in electronic health records-application of the unified theory of acceptance and use of technology. Appl Clin Inform 2020 May;11(3):415-426. [doi: 10.1055/s-0040-1712466] [Medline: 32521555]
32. Simons SM, Cillessen FH, Hazelzet JA. Determinants of a successful problem list to support the implementation of the problem-oriented medical record according to recent literature. BMC Med Inform Decis Mak 2016 Dec 02;16:102 [FREE Full text] [doi: 10.1186/s12911-016-0341-0] [Medline: 27485127]
33. Hersh WR. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care 2007 Jun;13(6 Part 1):277-278 [FREE Full text] [Medline: 17567224]
34. Classification suisse des interventions chirurgicales (CHOP) - Index systématique - Version 2021. Office Fédéral de la Statistique. 2020. URL: https://www.bfs.admin.ch/bfs/fr/home/actualites/quoi-de-neuf.assetdetail.13772935.html [accessed 2021-01-07]
35. Adaptation steps. Federal Institute for Drugs and Medical Devices. URL: https://www.dimdi.de/dynamic/en/classifications/icd/icd-10-gm/history/adaptation-steps/ [accessed 2020-11-25]
36. Kammermann M. Manuel de codage médical. Le manuel officiel des règles de codage en Suisse - Version 2020. Office Fédéral de la Statistique. 2019. URL: https://www.bfs.admin.ch/bfs/fr/home/actualites/quoi-de-neuf.assetdetail.9927930.html [accessed 2020-11-25]
37. SNOMED CT Starter Guide. URL: https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide [accessed 2021-01-07]
38. Agrawal A, He Z, Perl Y, Wei D, Halper M, Elhanan G, et al. The readiness of SNOMED problem list concepts for meaningful use of electronic health records. Artif Intell Med 2013 Jun;58(2):73-80. [doi: 10.1016/j.artmed.2013.03.008] [Medline: 23602702]
39. Galibois I, Academy of Nutrition Dietetics. Guide de Poche du Manuel de Référence de la Terminologie Internationale de Diététique et de Nutrition (TIDN) : Terminologie Normalisée Pour le Processus de Soins en Nutrition. Québec: Presses de l'Université Laval; May 2013:1-338.
40. Gaudet-Blavignac C. Geneva University Hospitals Common Problem List. Yareta. URL: https://doi.org/10.26037/YARETA:NAEGEJQVXZFWLIU236PXN5LUS4 [accessed 2022-07-20]

Abbreviations

CHOP: Swiss Classification for Surgical Interventions
EHR: electronic health record
HUG: Geneva University Hospitals
ICD-10: International Classification of Diseases, 10th revision
ICD-10 GM: International Classification of Diseases, 10th revision, German Modification
ICPC-2: International Classification of Primary Care, 2nd edition
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
WHO: World Health Organization

Edited by G Eysenbach; submitted 29.03.21; peer-reviewed by J Bettencourt-Silva; comments to author 20.04.21; revised version received 30.04.21; accepted 19.09.21; published 13.10.21.

Please cite as:
Gaudet-Blavignac C, Rudaz A, Lovis C
Building a Shared, Scalable, and Sustainable Source for the Problem-Oriented Medical Record: Developmental Study
JMIR Med Inform 2021;9(10):e29174
URL: https://medinform.jmir.org/2021/10/e29174
doi:10.2196/29174
PMID:34643542

Original Paper

A Patient-Screening Tool for Clinical Research Based on Electronic Health Records Using OpenEHR: Development Study

Mengyang Li[1,2], BSc; Hailing Cai[1,2], BA; Shan Nan[3], PhD; Jialin Li[1,2], MD; Xudong Lu[1,2], PhD; Huilong Duan[1,2], PhD

[1]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[2]Key Laboratory for Biomedical Engineering, Ministry of Education, Hangzhou, Zhejiang, China

[3]Hainan University School of Biomedical Engineering, Haikou City, China

Corresponding Author:
Xudong Lu, PhD
College of Biomedical Engineering and Instrument Science
Zhejiang University
Yuquan Campus
38 Zheda Road
Hangzhou, 310027
China
Phone: 86 13957118891
Email: lvxd@zju.edu.cn

Abstract

Background: The widespread adoption of electronic health records (EHRs) has facilitated the secondary use of EHR data for clinical research. However, screening eligible patients from EHRs is a challenging task. The concepts in eligibility criteria are not completely matched with EHRs, especially derived concepts. The lack of high-level expression of Structured Query Language (SQL) makes it difficult and time consuming to express them. The openEHR Expression Language (EL) as a domain-specific language based on clinical information models shows promise to represent complex eligibility criteria.

Objective: The study aims to develop a patient-screening tool based on EHRs for clinical research using openEHR to solve concept mismatch and improve query performance.

Methods: A patient-screening tool based on EHRs using openEHR was proposed. It uses the advantages of information models and EL in openEHR to provide high-level expressions and improve query performance. First, openEHR archetypes and templates were chosen to define concepts called simple concepts directly from EHRs. Second, openEHR EL was used to generate derived concepts by combining simple concepts and constraints. Third, a hierarchical index corresponding to archetypes in Elasticsearch (ES) was generated to improve query performance for subqueries and join queries related to the derived concepts. Finally, we realized a patient-screening tool for clinical research.

Results: In total, 500 sentences randomly selected from 4691 eligibility criteria in 389 clinical trials on stroke from the Chinese Clinical Trial Registry (ChiCTR) were evaluated. An openEHR-based clinical data repository (CDR) in a grade A tertiary hospital in China was considered as an experimental environment. Based on these, 589 medical concepts were found in the 500 sentences. Of them, 513 (87.1%) concepts could be represented, while the others could not be, because of a lack of information models and coarse-grained requirements. In addition, our case study on 6 queries demonstrated that our tool shows better query performance among 4 cases (66.67%).

Conclusions: We developed a patient-screening tool using openEHR. It not only helps solve concept mismatch but also improves query performance to reduce the burden on researchers. In addition, we demonstrated a promising solution for secondary use of EHR data using openEHR, which can be referenced by other researchers.

(JMIR Med Inform 2021;9(10):e33192) doi:10.2196/33192

KEYWORDS

openEHR; patient screening; electronic health record; clinical research

Introduction

Clinical research is a scientific research activity that considers patients as the main research object and focuses on the diagnosis, treatment, and prognosis of diseases. The identification of research subjects during clinical research is one of the major challenges. A study [1] in Britain showed that of the 114 surveyed clinical studies, only 35 (31%) could complete patient screening as planned. During the design of the research protocol, researchers develop detailed conditions for eligible patients. In the past, researchers collected eligible patients by asking clinicians or manually issuing recruitment ads. However, this is a labor-intensive and time-consuming task and can be helpful in small clinical research. The widespread adoption of electronic health records (EHRs) has enabled the secondary use of EHR data for clinical research.

However, there exist many obstacles to be overcome in using EHR data for clinical research. Fragmentation of clinical data and proprietary health information systems make it a challenge to adopt some specific screening methods [2-6]. These methods require detailed communication among researchers, clinicians, and information technology personnel each time. So, it is a time-consuming and error-prone process due to communication errors [7]. Only a few EHR vendors adopt health information standards and accommodate controlled terminologies [8]. Researchers have to express their query requirements into keywords to select patients from EHRs [9-12]. Due to these conditions, query tools based on EHRs are required for clinical research. Form-based query interfaces, such as Informatics for Integrating Biology & the Bedside (i2b2) [13], provide a promising direction for queries on EHRs. These interfaces partially meet query requirements by providing controlled query inputs for built-in coded concepts. However, complex screening conditions cannot be effectively and accurately expressed this way, especially for derived concepts. Wagholikar et al [14] proposed that derived concepts can only be expressed by Structured Query Language (SQL), which is a challenging task. The lack of domain-specific high-level expression of SQL makes it difficult for researchers to express these derived concepts. In addition, these query interfaces, such as i2b2, are mostly treated as clinical data warehousing and store EHR data through the star model. When faced with subqueries and join queries, query tools based on relational databases are inefficient [15].

Accordingly, a user-centered patient-screening tool with high-level expressions based on a standardized and scalable clinical data repository (CDR) can facilitate the use of EHR data in clinical research. OpenEHR is regarded as a promising tool to help build a CDR and support the expression of complex screening conditions. It provides a new formal modeling paradigm from clinical contents [16]. Its several features make it attractive in helping build patient-screening tools for clinical research. First, it provides open, semantically enabled, standard-based, vendor-independent, and use-case agnostic information models to represent clinical concepts [17]. It reuses existing archetypes in many particular clinical use cases across templates to reduce time and effort to enable semantic interoperability among different systems. This feature lays a solid foundation for the development of a CDR. Some studies on openEHR-based CDRs have been proposed [18-20]. Second, openEHR divides models into the archetype model (AM) and the reference model (RM). The AM can be used to represent domain knowledge. Within the AM, many coded values set or coding vocabularies can be drawn from controlled terminology resources [21-25]. Engineers only need to focus on developing software based on the RM, which facilitates maintainability [26]. This way, developers can provide different implementations for specific requirements. In addition, it provides openEHR openEHR Expression Language (EL) [27] to specify archetype rules and decision expressions. OpenEHR EL can be used to represent high-level query expressions combined with coding concepts. Domain-specific languages have shown promise in many use cases. So, openEHR EL makes it possible for clinical researchers to query on EHRs compared to SQL.

However, it is still an open question how openEHR can be applied in patient screening for clinical research. Specifically, two questions need to be tackled, the lack of high-level expressions and inefficient queries. Accordingly, in this study, openEHR EL is used to provide high-level expressions for queries, especially for derived concepts. Meanwhile, inefficient queries are generated for these derived concepts because they consist of simple concepts and complex constraints. Therefore, Elasticsearch (ES) [28] is introduced to build the underlying CDR for patient screening. By generating hierarchical indexes for corresponding archetypes and templates, our method avoids executing join queries and subqueries. To the best of our knowledge, there are almost no such query tools aimed at complex medical concepts in openEHR-based CDRs using ES.

The structure of the rest of this paper is as follows. Our method is proposed in the Materials and Methods section. After query requirements are collected, a representation method is proposed for eligible criteria based on archetypes and openEHR EL. Afterward, ES is used to generate hierarchical indexes based on archetypes. Finally, a screening tool is developed to support patient-screening tasks. The Results section gives the screening condition representation and execution performance evaluation. The Discussion section describes the contributions of our method and some relevant issues and future directions. Finally, conclusions are summarized.

Methods

Requirement Collection

Since it is difficult to collect actual query requirements in clinical research due to fragmented requests, conflicts of interest, security, etc, we considered clinical trials as representative examples of clinical research. A clinical trial is an experiment designed to answer specific questions about possible new treatments or new ways of using existing (known) treatments. To analyze the requirements of screening in clinical research, 389 stroke-related clinical trials were collected up to January 1, 2020, from the Chinese Clinical Trial Registry (ChiCTR) [29], including 2178 inclusion criteria and 2513 exclusion criteria, with a total of 4691 screening criteria. All these criteria were considered as the query requirements in this paper.

Representation of Screening Conditions

One of the major functions of patient-screening tools is to transform screening conditions in free text into computer-readable expressions. Ross et al [30] analyzed the composition and structure of screening conditions. Weng et al [31] surveyed the formal representation of eligibility criteria in clinical trials. Many representation methods in these studies can be used. One of the main considerations in the development of patient screening is to make it compatible between the representation of screening conditions and data representation in EHRs.

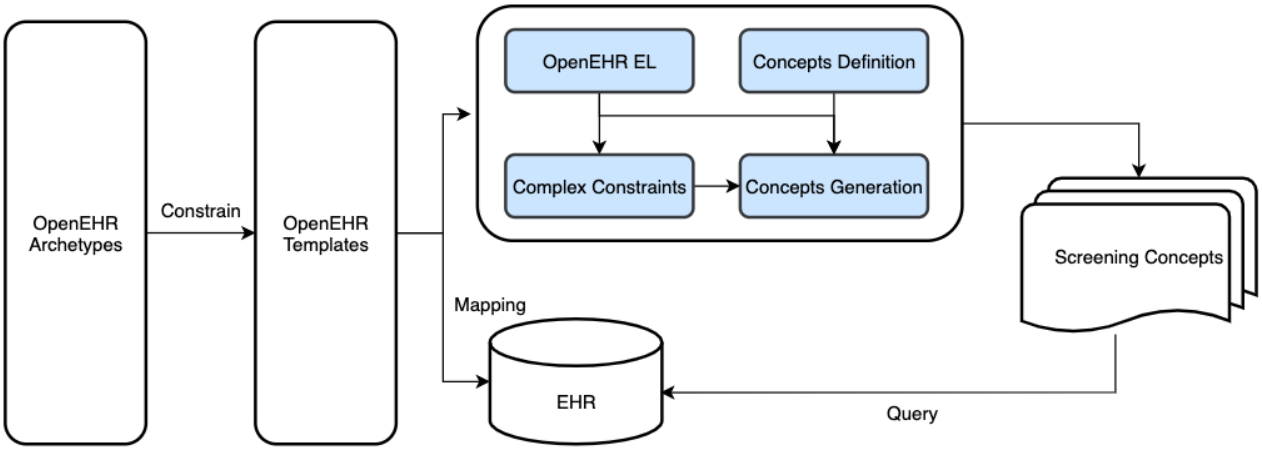
OpenEHR is proposed to represent the data structure in EHRs. EL is part of the openEHR specification for specifying archetype rules and decision language expressions. OpenEHR EL is based on the openEHR information model and is consistent with the structure in EHRs. Therefore, our study uses openEHR EL to represent screening conditions. OpenEHR EL provides complete arithmetic operators, relational operators, logical operators for different kinds of operations, and limited operations about time and collections. These operators do not meet the requirements of complex screening conditions in some cases (eg, for patients who meet the requirements of "white blood cell count continues to decrease in a specific duration after chemotherapy and radiation").

Consequently, for representing screening conditions, our method can be divided into two parts:

1. Defining concepts from openEHR archetypes and templates directly
2. Generating concepts by openEHR EL according to clinical requirements

The process of representing screening conditions is shown in Figure 1.

Figure 1. A process of representation for screening conditions. EHR: electronic health record; EL: Expression Language.



Screening Concept Definition

The screening conditions are composed of medical concepts and related constraints. Aiming at resolving the mismatch between screening concepts and data items in EHRs, a method was designed to define screening concepts from openEHR archetypes and templates, which are used to build data structures in EHRs. This way, screening conditions can be represented in a consistent way with EHRs to promote screening performance. For the management of screening concepts and the representation of complex constraints, relevant attributes need to be defined, as shown in Table 1.

Table 1. Definition of attributes of concepts.

Attribute	Description
Name	The name of a concept
Parent	Parent concept of the current concept, used to represent the hierarchical relationship among concepts
Path	The path of concepts as an identifier
Type	The data type of concepts, which decides allowed constraints
Unit	The unit of concepts, especially for laboratory test concepts
StartTime	The start time when an event happens, used to represent temporal constraints
EndTime	The end time when an event ends, used to represent temporal constraints
Value	The value of quantifiable concepts, such as the WBC[a] count, or used to represent some constants such as the lower limit of blood pressure

[a]WBC: white blood cell.

OpenEHR archetypes define complete domain contents for clinical concepts, which are loosely bound to attributes and constraints for the consideration of generality and reusability in design. The template is a reasonable composition of one or

several archetypes, which can be further constrained. Therefore, openEHR templates are used to directly define screening concepts. The rules for the definition of concepts are as follows:

1. A template corresponds to an archetype with detailed constraints, such as local optionality, default values. So, it is considered as a concept set. Each attribute node in the template is mapped to the subconcepts under the concept set.
2. When the attribute node in the template is mapped to a concept, the ontology name of the node is treated as the name of the concept. The data type of the node is used as the type of the concept, and the path of the node in the openEHR template is mapped to the path of the concept. For hierarchical relationships among different concepts, parent attributes can be identified by the preceding part of the children's paths because these paths can be treated as identifiers of these attributes.
3. For DV_QUANTITY attributes, the attribute "units" can be mapped to the unit attribute of the concept.
4. For DV_CODED_TEXT attributes, the attribute "defining_code" can be mapped to subconcepts under this concept. This attribute definition can be from different terminology services. Therefore, for the same subconcept, there may exist several data items under the DV_CODED_TEXT attribute's concept.
5. For other data types, relevant node items are just mapped to screening concepts according to step 2.

For the StartTime and EndTime of screening concepts, it is meaningless to extract from templates. Because these time attributes are related to specific clinical events, they can be specified to clinical concepts when screening patients. For example, the laboratory test may contain several concepts about time, such as result time, test time, and specimen receipt date and time. According to different application scenarios or the understanding of different researchers, the concept of a white blood cell (WBC) count can be bound to a different StartTime and EndTime.

Screening Concept Generation

Although some concepts can be defined according to rules directly from openEHR templates, there exist complex concepts and constraints that cannot be derived from these templates in practical cases, such as "white blood cell count continues to decrease in a specific duration after chemotherapy and radiation". For the constraint "decrease," no ready-made expression nor operators can represent it. In addition, chemotherapy is indicated by relevant drugs in general, such as altretamine, bendamustine, and azacytidine, instead of kept in EHRs directly. These expressions, such as chemotherapy, need to be specified by combining existing concepts and constraints.

OpenEHR EL is used to express complex constraints based on templates. Applying openEHR EL to generate customized concepts requires declaring variables, binding variables and data items in EHRs, and defining the logical expressions of simple concepts.

To declare variables, openEHR EL supports variable declarations, assignments, and expressions. So, customized concepts can be generated in the form of variables. For example, chemotherapy can be generated, as shown in Figure 2.

The mapping between openEHR templates and concepts has been realized in the process of screening concept definition and the structure of an EHR comes from archetypes and templates. As a result, the derived concepts are customized and can be directly used as the variable in EL expression to realize implicit binding.

Figure 2. The representation of "derived concepts" chemotherapy.

```
chemotherapy: List<String>
|
| an expression that will return true if chemotherapy
| contains specified values, like 'altretamine', 'bendamustine', 'azacitidine'
|
| these drugs can be replaced by other variables instead of literals.
|
chemotherapy.there_exists (
    agent (v: String): Boolean {
        v.contains('altretamine')
          or v.contains('bendamustine')
          or v.contains('azacitidine')
    }
)|
```

These customized concepts consist of logical expressions and defined concepts directly from templates. According to different types of constraints, the customized ways can be divided into three types:

1. The first way is to constrain defined concepts by arithmetic operators. For example, the body mass index (BMI) does not occur in EHRs directly. To screen patients by this condition, the BMI is defined as “BMI := weight/height^2”.
2. Another way is to generate customized concepts with relational operators. For example, cognitive impairment can be generated, as shown in Figure 3. The Mini-Mental State Exam (MMSE) is a widely used test of cognitive function. Any score of 24 or less (out of 30) indicates an abnormal cognition. This kind of knowledge can be an intuition for clinical researchers on stroke but is not identified by screening tools.
3. Complex customized concepts can be generated by combining arithmetic operators or condition chains by logical operators. Figure 4 gives an example of this. The cognitive impairment diagnosis concept can be defined directly from templates about problem/diagnosis. Two test scales are used to measure the different levels of cognitive impairment. By combining these three expressions, the new customized concept can help screen more eligible patients with high accuracy.

Figure 3. The representation of derived concepts, which is based on relational operators.

```
| It's a 30-point questionnaire and is used in clinical research to measure cognitive impairment
mmse:= mini_mental_state_examination
| according to the scores, cognitive impairment can be represented.
cognitive_impairment:= mmse < 24
```

Figure 4. The representation of derived concepts by logical operators.

```
cognitive_impairment_diagnosis:= problem_diagnosis = "cognitive impairment"

moca:= montreal_cognitive_assessment
mmse:= mini_mental_state_examination
cognitive_impairment:= cognitive_impairment_diagnosis or mmse < 24 or moca < 26
```

There are other issues to consider in the generation of customized concepts. One of the issues is the generation of nested concepts. Because some concepts are complex, only simple concepts defined from templates cannot meet the requirements of clinical scenarios. In these cases, intermediate concepts need to be first generated, and then customized concepts are expressed based on these intermediate concepts. For example, although obesity/overweight can be defined in openEHR templates with a specific terminology from the International Classification of Diseases, Tenth Revision (ICD-10), numbered E66, some expressions can be treated as the same criteria semantically, as follows in Figure 5.

Figure 5. The representation of intermediate concepts for derived concepts.

```
| here, we treat people who are not less than 18 years old as adults.
adults:= age > 18

| define body mass index as an intermediate concept
BMI:= weight/height^2

| define obesity from diagnosis
obesity_diagnosis:= problem_diagnosis = "obesity" or problem_diagnosis = "overweight"

| define obesity from terminology
icd_10_obesity := problem_diagnosis.code = "E66"

| define obesity according to BMI
bmi_obesity := adults and BMI > 30

| define final obesity concept
obesity := obesity_diagnosis or icd_10_obesity or bmi_obesity
```

Constraints About Screening Concepts

Operators provided by openEHR EL meet the maximum requirements of representation to express existential, arithmetic, logical, and relational constraints. In some cases, it is a challenge to represent constraints about collections and time.

The screening conditions about collections are highly flexible. In general, knowledge engineers are required to define these constraints for specific clinical requirements. For different medical institutions, even different clinicians, the understanding of these constraints can be different. Therefore, according to collected screening conditions, some constraints are predefined for convenience, as shown in Table 2.

Table 2. Predefined constraints for collections.

Constraint name	EL[a] expression	Parameter
First time	[concept].count()=1	Concept name
Stable	[concept].max()-[concept].min()<[value]	Concept name; self-defined threshold value for the comparison
Increase	[concept].last()>[concept].first()	Concept name
Decrease	[concept].last()<[concept].first()	Concept name

[a]EL: Expression Language.

To further constrain some concept set, the screening tool supports clinical research to self-define new constraints and edit existing constraints, in addition to using predefined constraints.

Allen et al [32] summarized 13 temporal representation patterns for comparing two events. They are represented by expressions in our tool, as shown in Table 3.

These temporal constraints cannot meet the requirements in some cases, for example, “The patient was treated with heparin within 48 hours.” According to the analysis about collected criteria, an “interval” attribute was introduced into our tool based on Allen et al's [32] patterns to represent the interval between different clinical events, for example, “diff([concept1].StartTime, [concept2].EndTime) [><=] Interval”.

By combining these extended constraints and the 13 patterns proposed by Allen et al [32], most screening conditions can be represented about temporal constraints. For example, “The patient was treated with heparin within 48 hours” can be represented in two ways, as shown in Figure 6.

Table 3. Representation of Allen temporal patterns.

Name	Expression	Description
Before	[concept1].StartTime<[concept1].EndTime<[concept2].StartTime<[concept2].EndTime	Concept1 occurs before Concept2.
Meets	[concept1].StartTime<[concept1].EndTime=[concept2].StartTime<[concept2].EndTime	Concept2 occurs at the end of Concept1.
Overlaps	[concept1].StartTime<[concept2].StartTime<[concept1].EndTime<[concept2].EndTime	Concept1 occurs before Concept2 and ends before Concept2.
Begins	[concept1].StartTime=[concept2].StartTime<[concept1].EndTime<[concept2].EndTime	Concept1 and Concept2 occur at the same time, and Concept1 ends first.
BegunBy	[concept1].StartTime=[concept2].StartTime<[concept2].EndTime<[concept1].EndTime	Concept1 and Concept2 occur at the same time, and Concept2 ends first.
During	[concept2].StartTime<[concept1].StartTime<[concept1].EndTime<[concept2].EndTime	Concept1 occurs after Concept2, and Concept1 ends before Concept2.
Contains	[concept1].StartTime<[concept2].StartTime<[concept2].EndTime<[concept1].EndTime	Concept1 occurs before Concept2, and Concept1 ends after Concept2.
Equals	[concept1].StartTime=[concept2].StartTime<[concept2].EndTime=[concept1].EndTime	Concept1 and Concept2 occur and end at the same time.
OverlappedBy	[concept2].StartTime<[concept1].StartTime<[concept2].EndTime<[concept1].EndTime	Concept1 occurs after Concept2, and Concept1 ends after Concept2.
Ends	[concept2].StartTime<[concept1].StartTime<[concept1].EndTime=[concept2].EndTime	Concept1 occurs after Concept2, and both end at the same time.
EndedBy	[concept1].StartTime<[concept2].StartTime<[concept2].EndTime=[concept1].EndTime	Concept1 occurs before Concept2, and both end at the same time.
MetBy	[concept2].StartTime<[concept2].EndTime=[concept1].StartTime<[concept1].EndTime	Concept1 occurs at the end of Concept2.
After	[concept2].StartTime<[concept2].EndTime<[concept1].StartTime<[concept1].EndTime	Concept2 occurs before Concept1.

Figure 6. An example of temporal constraints.

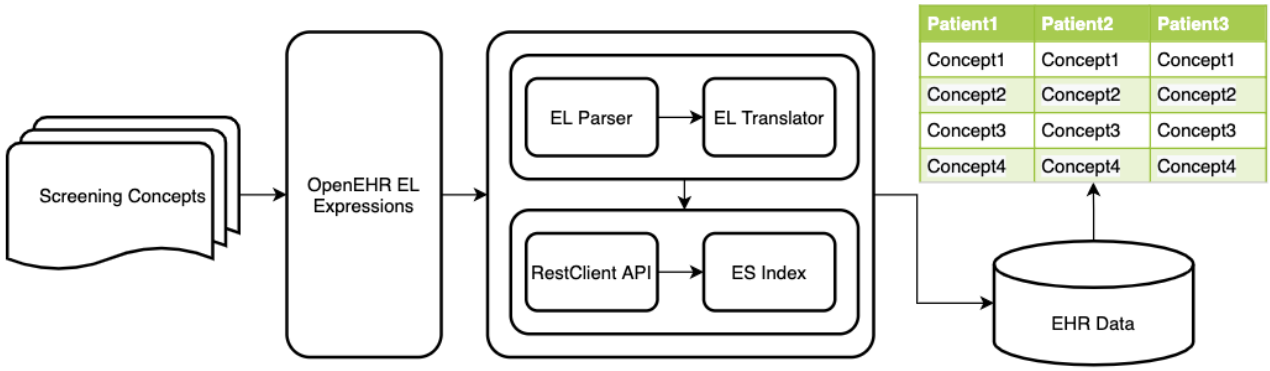
```
| ----- Method 1 -----
| a virtual concept is meaningless clinically and used to compare
| with other concepts purely.
| here, the virtual concept represents now.
virtual_concept.StartTime := {Env}.current_date_time()
virtual_concept.EndTime := {Env}.current_date_time()

| define heparin treatment
heparin_drug := drug_name = "heparin"
heparin_drug.StartTime := medication_Order_start_time
heparin_drug.EndTime := medication_Order_stop_time

results := attached(heparin_drug) and diff_duration(virtual_concept, heparin_drug) < P48H

| ----- Method 2 -----
| define heparin treatment
heparin_drug := drug_name = "heparin"
heparin_drug.StartTime := medication_Order_start_time
heparin_drug.StartTime := medication_Order_stop_time

results := attached(heparin_drug) and diff({Env}.current_date_time(), heparin_drug.StartTime) < P48H
```

Execution of Screening Conditions

Different from the method where the screening expression and constraints are directly hard-coded, such as i2b2, ANother Tool for Language Recognition (ANTLR) [33] is used to parse openEHR EL expressions. In addition, a translator is implemented to transform the screening conditions into the underlying query language. The decoupling design uses the mechanism of openEHR two-level modeling and makes it easy to keep maintainability.

Most openEHR-based CDRs are based on relational databases [19,20]. The data scattered in multiple tables make it unavoidable to bring multitable joins and subqueries. To improve query performance, this study decided to use a dedicated search engine to execute queries. ES can store, search, and analyze a large amount of data in a short time and can meet the performance requirements of patient screening as a distributed search engine. Similar to relational databases, each typed field can be indexed and queried. The architecture is designed for the execution of queries in Figure 7.

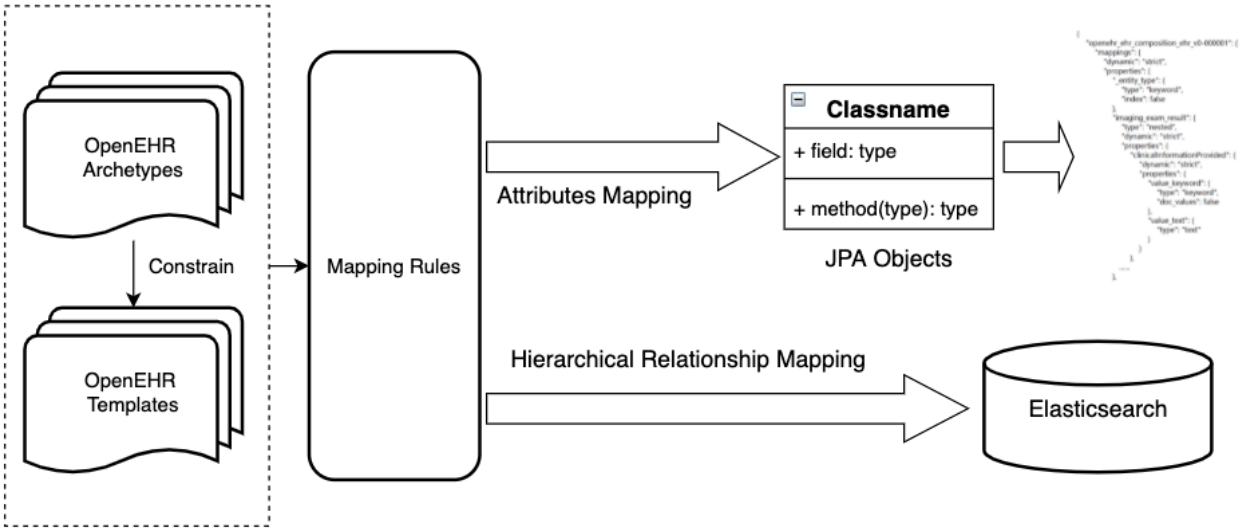
Figure 7. The architecture of query execution for screening conditions. API: Application Programming Interface; EHR: electronic health record; EL: Expression Language; ES: Elasticsearch.



Data are stored as a document in ES, and the document's schema is defined by mapping. The index is similar to a database, and the type is similar to the table structure in a relational database. ES documents store data in the form of key–value pairs, and documents can also be used as class types of values to achieve hierarchical storage. To query EHR data in an index, this study proposed a method that maps openEHR templates to index structures.

Specifically, openEHR templates are used to generate corresponding entities conformed to the Java Persistence Application Programming Interface (API), or JPA [34], according to a series of rules. Then, mapping relationships between entities and indexes are described according to the ES-related annotations provided by a hibernate search [35]. Finally, a schema of the index structures is generated by using the existing hibernate search framework. The flowchart is shown in Figure 8.

Figure 8. The flowchart of mapping between templates and ES. API: Application Programming Interface; EHR: electronic health record; ES: Elasticsearch; JPA: Java Persistence API.



Mapping Rule Definition

Mapping rules decide the detailed structure of the index, which plays a significant role in performance. Template index-mapping rules can be divided into several parts:

1. Data type mapping
2. Hierarchical relationship mapping
3. Naming strategies

A composition archetype and a corresponding template are required as a container to import other archetypes, including demographic, imaging examination, laboratory test, problem diagnosis, medication order, and procedure. The composition template is designed into a single index in ES. Corresponding to the definition of screening conditions, attribute nodes and the hierarchical relationship are the focus in the mapping process.

Data Type Mapping

The data types of openEHR template attributes are defined in the RM, and since each field of ES has a type, each RM needs to add corresponding field-type annotations to map to the corresponding entity object types. Table 4 shows the mapping relationships between commonly used data types and entities.

Field annotations decide not only the data types in ES but also the analyzers that carry out indexing and text processing used in these fields. All data types are mapped into three field types:

1. GenericField: Use a default field type and analyzer for the specific attribute type according to provided strategies by the hibernate search.
2. KeywordField: It only works for string fields whose value is treated as a single keyword. So, it is appropriate for terminology-constrained diagnosis, laboratory tests, imaging examination, etc.
3. FullTextField: Compared with KeywordField, data are treated as free text, and so this only works for string fields. The text is split into several tokens as an index. Here, for string fields in DV_CODED_TEXT and DV_TEXT, two kinds of field annotations are added so that they can be queried in different ways.

Table 4. Data type mapping rules for openEHR[a].

Data type	Attribute	Field type	Field annotation
DV_BOOLEAN	value	Boolean	GenericField
DV_CODED_TEXT	value	String	KeywordField; FullTextField
	code	String	KeywordField
DV_COUNT	magnitude	Integer	GenericField
DV_DATE	dateTime	LocalDate	GenericField
DV_DATE_TIME	dateTime	LocalDate	GenericField
DV_DURATION	duration	Duration	GenericField
DV_IDENTIFIER	id	String	KeywordField
DV_QUANTITY	magnitude	Double	GenericField
	units	String	KeywordField
DV_TEXT	value	String	KeywordField; FullTextField
DV_URI	uri	URI	GenericField

[a]EHR: electronic health record.

Hierarchical Relationship Mapping

The composition template corresponds to only one index structure. The imported archetypes generate hierarchical relationships in this index. Attributes with basic data types, such as string, int, and date, can be indexed as value fields. These imported archetypes and underlying slotted archetypes and attributes of collection types are indexed as object fields. Several Java Persistence entities are built according to these archetypes. The detailed rules are as follows:

1. Each archetype corresponds to a master entity object that contains the indexed annotation.
2. Slotted archetype and collection-type attributes are mapped into a new embedded entity in the master entity. However, there exists a difference between these two types. A slotted archetype can be treated as an independent medical concept and is considered as a normal object field, which means all of its attributes are flatted during indexing. For an attribute of collection type, it is mostly attached to a major medical concept and cannot be used on its own. Therefore, it is mapped into a nested field type, which can keep the original structure in the archetypes.
3. Every generated entity is labeled with an ID field that uniquely identifies a document.
4. Every generated entity is also labeled with a timestamp field, which is treated as a major time marker for comparison, which can be specified during data integration.
5. Every field type defined in the openEHR RM is considered as an embedded entity.

Naming Strategies

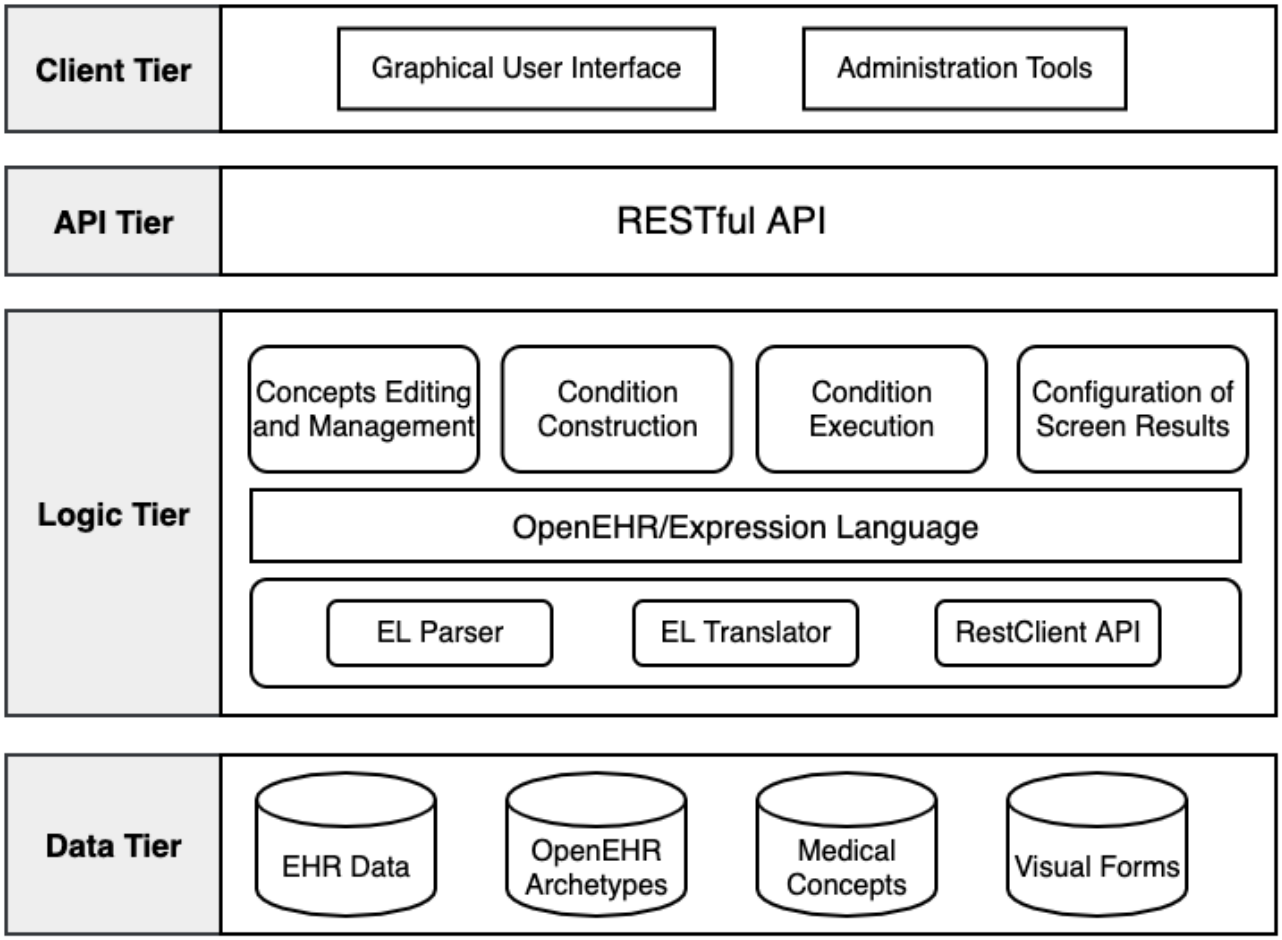
1. The name of attributes in archetype data types is predefined according to Java field names. Particularly, the Java string field, which is annotated with two annotations, KeywordField and FullTextField, is mapped to two fields named "{java field name}_keyword" and "{java field name}_text".
2. For attributes within entry archetypes directly, the names are defined with their textual name mentioned in the ontology section of archetypes. This text is joined with "_". For unique identification in the index, every attribute's name is prefixed with the archetype concept name; for example, "Body site" in openEHR-EHR-EVALUATION.problem_diagnosis.v1 will be named with "problem_diagnosis_Body_site".
3. For attributes within slotted and collection-type archetypes, the name is decided by two parts: one is these direct parent archetypes; the other is entry archetypes imported in the composition template. For multiple levels of archetypes, the name is provided recursively.

Development of a Screening Tool

Based on the proposed method above and requirements analysis, we developed a patient-screening tool on EHRs using openEHR. The system architecture is shown in Figure 9. Our tool uses a loosely coupled architecture where all modules are connected loosely so that they can be maintained and replaced more easily. To be specific, the system is mainly divided into three parts:

1. Concept editing and management: This is to realize the maintenance and management of screening concepts by definition and generation. Clinical researchers can edit and revise these concepts according to specific requirements.
2. Screening conditions' construction/execution: An easy-to-operate visual interface is provided for users to edit screening conditions, and then restful APIs are used to execute queries in ES.
3. Results of screening configuration: Aimed at different data requirements, researchers can predefine specific data views in forms. This module makes it more convenient to get access to screening results by customized views.

Figure 9. System architecture for the patient-screening tool. API: Application Programming Interface; EHR: electronic health record; EL: Expression Language.

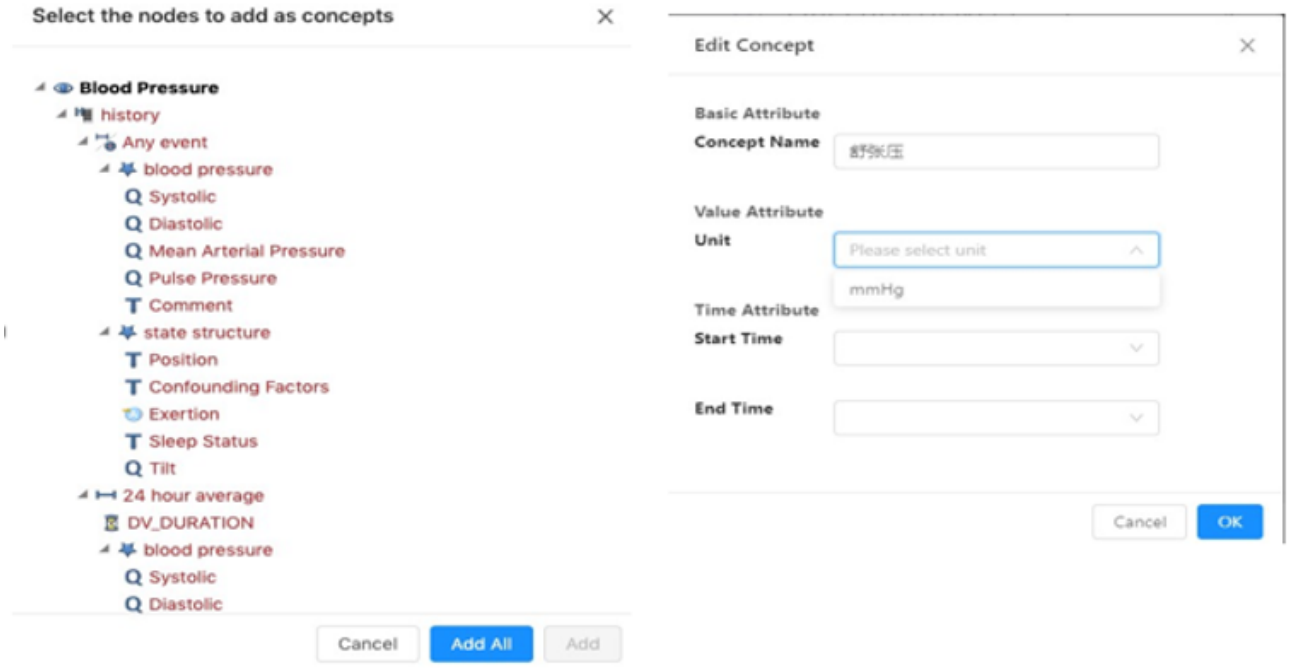


Concept Editing and Management

The main function of this module is to manage screening concepts and to provide a concept generation function based on openEHR templates and openEHR EL so that users can quickly realize the mapping between screening concepts and EHR data.

OpenEHR templates used for EHRs are obtained from the template repository and parsed by tools provided by the openEHR community [36]. The screening concepts are generated based on the obtained templates. Since templates used by the electronic medical record system are for routine delivery of health care, part of them is not required for clinical research. So, these templates and related nodes can be selected according to real situations. Basic concepts can be defined from these templates, complex or derived concepts should be generated by the basic concepts, and constraints should be provided by openEHR EL, as shown in Figure 10.

Figure 10. Editing concepts after selection in template.



Screening Conditions' Construction/Execution

Screening conditions' construction/execution is the core component of this screening tool, with which researchers formally construct screening conditions. Based on the screening conditions, a query is executed by calling the restful API provided by ES.

A graphical editing interface is provided to support the construction of screening conditions. It is designed to organize components in hierarchical form to accomplish construction. Screening conditions can be divided into different groups corresponding to different visual components. Every screening concept is filled into a single group. Meaningful feedback is necessary because researchers do not know the exact data in EHRs. When constructing screening conditions, the number of results for the screening conditions in every group is immediately queried to support the revision of conditions. By default, groups are connected by logic conjunctions. Logical disjunction can be used within groups. For a single group for a screening concept, multiple constraints can be added by just dragging visual components filled with necessary information. Temporal constraint controls can be added between different groups. The user interface is shown in Figure 11.

Screening conditions contain complex collection and temporal constraints. Users can define customized visual components of constraints according to their own needs. Through the defined visual components, the corresponding openEHR EL expressions are generated to express complex conditions, as shown in Figure 12.

Screening conditions are constructed each time from scratch. It is unnecessary and time consuming because some conditions seem to be similar to a certain extent, such as conditions including the same concepts. So, our tool provides a function for saving constructed screening conditions for reuse after execution. Later, users can reuse the screening conditions without construction again. These screening conditions are saved in hierarchical form, and the required screening conditions are dragged and dropped at the corresponding level of the current conditions.

Figure 11. The user interface for construction of screening conditions.

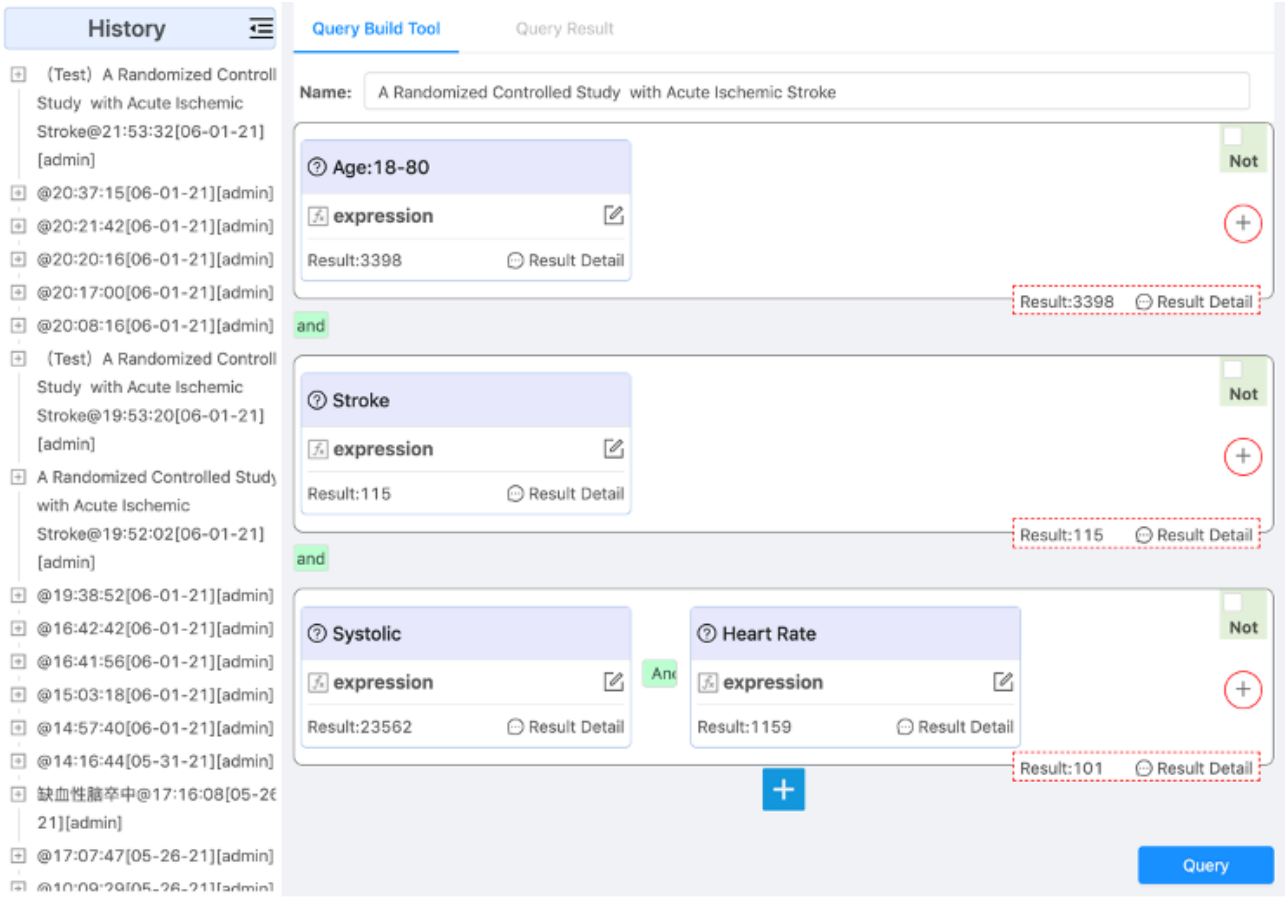


Figure 12. Derived concepts' generation.

Results of Screening Configuration

The screening results' viewing assists researchers to view the details of selected patients to further determine whether the patients meet the requirements. Some research platforms usually use the case report form (CRF) to record patient data. Our tool relies on a developed CRF generation service to help researchers customize patient result forms so that they can view information according to specific needs.

Results

Experiments for Representation

An experiment was carried out in an openEHR CDR based on a grade A tertiary hospital in China. In total, 500 sentences in the collected 389 clinical trials were randomly selected to analyze and evaluate the effectiveness of the proposed tool. A clinician and two information technology personnel joined the experiment. The clinician was responsible for providing medical domain knowledge and giving reliable proof when faced with different opinions. One of the information technology personnel took charge of the issues about openEHR, and the other was the core developer of the tool.

In this CDR, more than 30,000 concepts were directly defined from 34 openEHR templates. These concepts were used to represent 500 eligible criteria mentioned above. In these conditions, 589 concepts were found. Among the concepts found, 513 (87.1%) concepts could be represented and 471 concepts could be directly defined from templates. In addition, 42 concepts needed to be generated by the configurable operation, and there still existed a challenge to represent 76 concepts. Table 5 shows the part of configurated concepts.

At the same time, our experiment was also carried out in an i2b2 web client to figure out the differences between our tool and the client. By comparing the provided query functions, the differences are given in Table 6 (Y means yes, and N means no).

Table 5. Part of configurated concepts.

Name	Descriptive expression	Mentioned times (n)
Course of a disease	current_date_time() - encounter.StartTime	78
Stable condition	(Systolic blood pressure≥120 and systolic blood pressure≤220) and (40≤heart rate≤100) and blood oxygen saturation≥92% and body temperature≤38.5	10
In good spirits	27<mmse[a]<30	2
Dual-antiplatelet therapy	attached(aspirin) or attached(clopidogrel)	1
Cognition impairment	cognitive_impairment_diagnosis or mmse<24 or moca[b]<26	11
Obesity/overweight	obesity_diagnosis or icd_10_obesity or bmi_obesity	2
Psychotropic drugs	attached(sulpiride) or attached(risperidone)	2

[a]mmse: Mini-Mental State Exam.

[b]moca: Montreal Cognitive Assessment.

Table 6. Comparison with the i2b2[a] web client.

Constraint support	Details	Our tool	i2b2	Example
Exist	—	Y	Y	Patients with stroke
Relational	—	Y	Y	Patients aged 30-80 years
Logical	—	Y	Y	Cerebral infarction or cerebral hemorrhage and subarachnoid hemorrhage
Temporal	Duration	Y	N	Operation time <1 h or >3 h
	Interval among different clinical events	Y	Y	Antiplatelet drugs within 2 weeks before surgery
	Interval related to a single clinical event	Y	N	Treated with botulinum toxin injection within 6 months
Collection	Count	Y	Y	First onset
	Complex computation on collections	N	N	Average treatment time less than 1 h per week
Self-defined constraint	—	Y	N	WBC[b] count decreasing
Self-defined concept	—	Y	N	Stable condition

[a]i2b2: Informatics for Integrating Biology & the Bedside.

[b]WBC: white blood cell.

Evaluation for Performance

An evaluation was performed to validate the query performance of our method. The used data are from the hospital mentioned before. EHRs store all the information generated during the routine delivery of health care, and part of them is about management and charge information, which are not required by clinical research. So, a total of seven archetypes and related templates were selected, including demographics, examinations, and laboratory tests. The selected templates are shown in Table 7. Considering the sensitivity of data and the complexity of data integration, only data related to cerebrovascular diseases were extracted, including 95,226 records of demographics, 449,880 records of admission, 4,239,454 records of physical sign information, 5,832,990 records of laboratory tests, 12,966,659 records of order information, 158,240 records of diagnostic information, and 176,798 records of imaging examination. The data storage structure was generated based on archetype relational mapping [19] and our template index-mapping method and is shown in Figure 13 and Figure 14, respectively.

Table 7. OpenEHR[a] archetypes for EHR structures.

Name	Archetype
Person	openEHR-EHR-ADMIN_ENTRY.person.v1
Patient admission	openEHR-EHR-ADMIN_ENTRY.Patient_Admission.v2
Laboratory test	openEHR-EHR-OBSERVATION.lab_test_single.v1
Order	openEHR-EHR-INSTRUCTION.order.v1
Imaging examination	openEHR-EHR-OBSERVATION.Imaging_examination_report.v2
Physical sign	openEHR-EHR-OBSERVATION.physical_sign.v1
Problem diagnosis	openEHR-EHR-EVALUATION.problem_diagnosis.v1

[a]EHR: electronic health record.

Figure 13. Database schemas by archetype relational mapping.

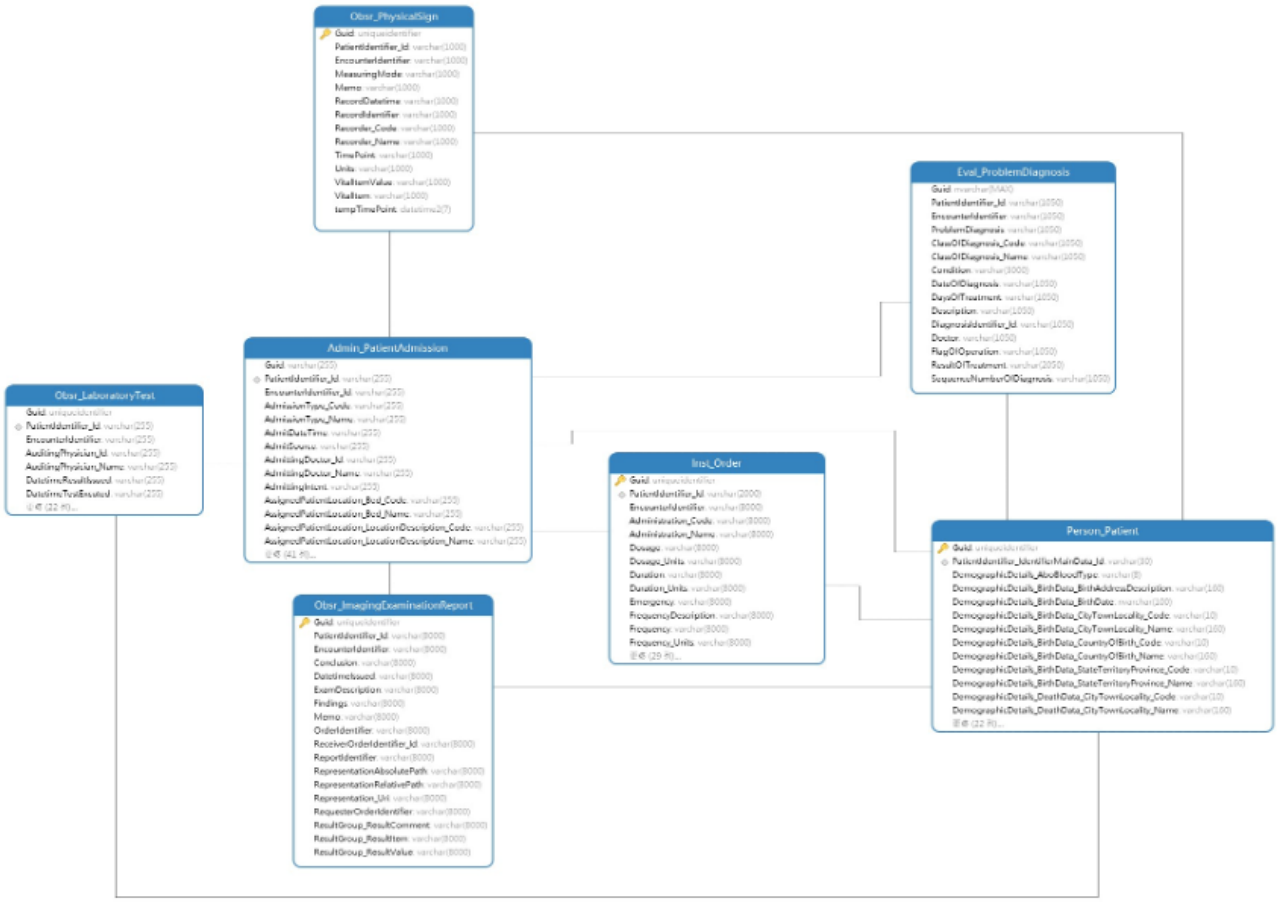


Figure 14. Index structure for the openEHR template. EHR: electronic health record.

```
{
  "openehr_ehr_composition_ehr_v0-000001": {
    "mappings": {
      "dynamic": "strict",
      "properties": {
        "_entity_type": {
          "type": "keyword",
          "index": false
        },
        "imaging_exam_result": {
          "type": "nested",
          "dynamic": "strict",
          "properties": {
            "clinicalInformationProvided": {
              "dynamic": "strict",
              "properties": {
                "value_keyword": {
                  "type": "keyword",
                  "doc_values": false
                },
                "value_text": {
                  "type": "text"
                }
              }
            },
            ......
        },
        "lab_test_single": {
          ......
        },
        "order": {
          ......
        },
        "patient_Admission": {
          ......
        },
        "person": {
          ......
        },
        "physical_sign": {
          ......
      }
    }
  }
}
```

The test cases were executed in Windows 10 with 16GB RAM and an Inter(R) Core (TM) i5-4590 CPU including Microsoft SQL Server 2014-12.0.2269.0 and Elasticsearch-7.11.1.

Our study selected six screening conditions as test cases, and each test case was tested five times in two test environments to eliminate accidental errors. The execution time of the screening conditions was separated into two parts: translation time of expressions and query time in underlying persistence layers. The selected six test cases are shown in Table 8.

The test results of these test cases are shown in Table 9 (for archetype relational mapping) and Table 10 (for template index mapping). Their comparison result is shown in Figure 15.

Table 8. Test cases for performance evaluation.

Test case	Condition	Description
Query1	Patients with evacuation of intracerebral hematoma	Occur in a single table, no join operation required
Query2	Female patients between 20 and 60 years	Occur in a single table, no join operation required
Query3	60-70-year-old female patients diagnosed with cerebral hemorrhage or cerebral infarction	Occur in two table, join operation between two tables required
Query4	Women between 60 and 70 years diagnosed with cerebral hemorrhage or cerebral infarction taking aspirin	Occur in three tables, join among three tables required
Query5	Female patients between 60 and 70 years diagnosed with cerebral hemorrhage or cerebral infarction undergoing a WBC[a] laboratory test and taking aspirin	Occur in four table, join operation among four tables required
Query6	The last WBC count more than 10×10^9/L	Sort and aggregation operation probably required

[a]WBC: white blood cell.

Table 9. Test results for archetype relational mapping.

ID	SQL[a] execution time without index (ms)	SQL execution time with index (ms)	Number of results (n patients)
Query1	1547	30	40
Query2	256	59	8570
Query3	284	257	1536
Query4	5253	183	154
Query5	6497	4893	106
Query6	2484	2193	14,583

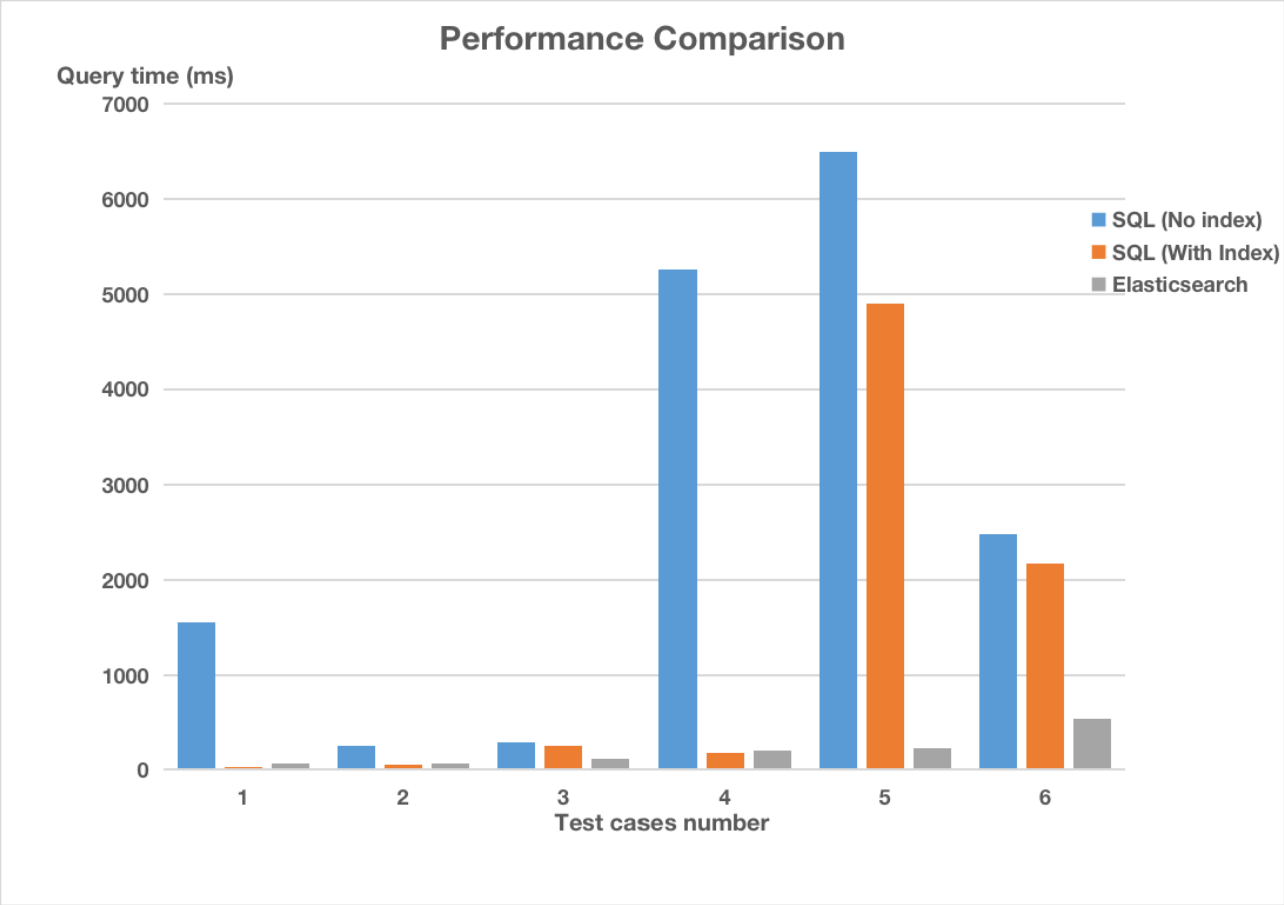
[a]SQL: Structured Query Language.

Table 10. Test results for template index mapping.

ID	Translation time of EL[a] (ms)	Query time (ms)	Total time (ms)	Number of results (n patients)
Query1	46	22	68	40
Query2	62	10	72	8570
Query3	62	49	111	1536
Query4	136	64	200	154
Query5	153	75	228	106
Query6	169	364	533	14,583

[a]EL: Expression Language.

Figure 15. The performance comparison results between two methods. SQL: Structured Query Language.



The test results show that the number of screening results obtained by the two methods is the same, and it can be considered that the screening tool proposed can correctly obtain screening results.

Considering query performance, in general, SQL queries with an index outperform SQL queries without an index. For queries using SQL (Query1 and Query2) without joins, the more data in the table, the longer the execution time. In addition, the query time increases rapidly (Query3, Query4, and Query5) during joins among three tables or even four tables.

With regard to execution using ES, it does not show obvious advantages among queries without joins, such as Query1 and Query2. However, for queries that are related to joins, ES outperforms SQL queries because the patient data are indexed with a single document, there is no need for joins among different documents, and the performance is more stable. At the same time, there exists only a small gap between the execution time of the search engine and the translation time of screening conditions. Considering the volume of EHR data on disks and the translation running in memory, the translation time has a serious influence on query performance. Therefore, there is still significant room for further improvement.

In comparison between the two methods, the screening execution method in our study showed better performance than the SQL-based method, especially in the screening context of multitable joins (Query4 and Query5).

Discussion

Principal Results

We proposed a patient-screening tool using openEHR to transform screening conditions into expressions for queries on EHRs. The tool is designed to support queries on EHRs directly within a local context. To sum up, our tool has the following features:

- First, the tool supports definition and generation from openEHR archetypes and templates. These concepts can be simple concepts and derived concepts. In previous studies, many tools have just provided a fixed-concept set based on some terminologies. Although a related study is proposed to extend concepts, these extended concepts can be only added by SQL expressions, which is a big challenge for clinical researchers.
- Second, the tool improves the performance of screening compared with SQL-based methods. With the continuously increasing data in EHRs, there is a serious bottleneck for queries in these situations. Our method proposes an implementation of openEHR AMs to promote query enhancement based on ES. It is worth taking as a reference to design other query tools not limited to clinical research.
- Third, the tool provides a promising solution for secondary use of EHRs in clinical research for the openEHR community. To the best of our knowledge, there is no such tool based on openEHR. Our study shows that although openEHR specifications are mostly designed for the EHR

environment, they can be used for clinical research in a way proposed in this study. Although our method is proposed within the context of openEHR, other information models can be translated into openEHR information models, which is proved by previous research [37]. In this way, our method can be used in these information models.

Ability of Representing Complex Concepts

According to the results of experiments for representation, 87.1% of concepts can be represented by our method. In addition, 76 concepts (12.9%) were not expressed successfully because of the complexity of screening conditions. Some reasons are as follows:

- First, no appropriate archetypes or templates can be used to generate these concepts. In other words, the necessary basic concepts are not covered by the EHR information models. In addition, the concepts that occurred in conditions are not recognizable by our templates due to differences. With this kind of issue, more archetypes and templates need to be encouraged to be developed for specific requirements. In addition, more local knowledge should be introduced into templates.
- Second, some concepts do not occur in a structured way. For example, “severe coronary stenosis” is mostly recorded in the description of imaging examination. In these situations, concepts cannot be defined from openEHR templates directly for screening. This limitation can be solved in two ways. One is to do what i2b2 does. New concepts can be extracted from medical texts with natural language processing (NLP), and a mapping relationship can be built between these new concepts and original text for backtrack during querying. Another way is to process these medical texts independently, and new strategies can be proposed to query texts together with structured data.
- Third, some concepts mentioned in these conditions are coarse grained and fuzzy. It is difficult to define a comprehensive expression to meet all requirements for all queries because of different knowledge backgrounds and considerations. For example, in the condition “patients diagnosed with the diseases that may lead to dysphagia,” it is difficult to represent ”the diseases.“ The definition of ”the diseases“ is general involving many concepts. Diseases leading to dysphagia can be of different types, including brain/nervous system diseases, muscle diseases, and esophageal diseases. The results of the definition can be a big list. In addition, different departments focus on different types of diseases. For example, stroke, Parkinson's disease, and other brain/nervous system diseases are considered in psychiatry departments. Oncology departments tend to consider esophageal carcinoma when finding dysphagia. The general description of concepts in screening conditions is common, and they play a significant role in hindering accurate querying.

Limitations and Future Directions

Visual editing tools can reduce the burden of researchers, but they still require a certain amount of manual work. For example, such tools provide some modules such as reusing existing conditions for convenience. Developing new screening conditions from scratch is still inevitable, especially for complex conditions and conditions that have not been used before. With the continuous development of NLP technology, screening conditions in free text can be automatically converted into executable queries, such as SQL or openEHR EL expressions in this paper.

Stubbs et al [38] proposed the task of patient screening using EHRs. Some contestants [39,40] used text data to determine which patients meet the criteria by using rule-based methods, neural networks, etc. Some studies [41-44] have transformed conditions into a computer-interpretable format with information extraction. Criteria2Query [45] provides a natural language interface to help find eligible patients. In addition, some text-to-SQL methods are proposed to execute queries on EHRs [46]. This reduces the workload by predicting the SQL query for a given condition about a database.

To some extent, these studies greatly relieve the burden of researchers by allowing familiar way of humans to construct queries. Their work reduces the extensive interaction issues with systems/databases or administrators. However, the end-to-end process hinders manual participation, considering extremely complicated conditions. Meanwhile, for concept mismatch between conditions and EHR data, they do not provide an available solution.

Considering the advantages of NLP technology and the method proposed in this paper, the future direction for us is to combine machine learning methods, rule-based methods, and engineering to improve queries on EHR. Their combination will outperform any single method.

Conclusions

In this paper, we developed a patient-screening tool for clinical research using openEHR. The tool helps solve concept mismatch, especially for derived concepts. The use of ES improves query performance compared with SQL-based methods. The tool is applied to stroke-related clinical research and shows promise. Moreover, we demonstrated a promising solution for secondary use of EHR data using openEHR. In the future, we will enhance the tool by leveraging NLP techniques to enable automatic query formulation for simple and derived concepts to further reduce the burden of researchers.

Acknowledgments

This study was funded by the Chinese National Science and Technology Major Project (grant nos. 2020YFC2003401 and 2016YFC0901703).

Conflicts of Interest

None declared.

References

1. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials 2006 Apr 07;7:9 [FREE Full text] [doi: 10.1186/1745-6215-7-9] [Medline: 16603070]
2. Embi PJ, Jain A, Clark J, Harris CM. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc 2005:231-235 [FREE Full text] [Medline: 16779036]
3. Ahmad F, Gupta R, Kurz M. Real time electronic patient study enrollment system in emergency room. AMIA Annu Symp Proc 2005:881 [FREE Full text] [Medline: 16779168]
4. Beauharnais CC, Larkin ME, Zai AH, Boykin EC, Luttrell J, Wexler DJ. Efficacy and cost-effectiveness of an automated screening algorithm in an inpatient clinical trial. Clin Trials 2012 Apr;9(2):198-203 [FREE Full text] [doi: 10.1177/1740774511434844] [Medline: 22308560]
5. Heinemann S, Thüring S, Wedeken S, Schäfer T, Scheidt-Nave C, Ketterer M, et al. A clinical trial alert tool to recruit large patient samples and assess selection bias in general practice research. BMC Med Res Methodol 2011 Feb 15;11:16 [FREE Full text] [doi: 10.1186/1471-2288-11-16] [Medline: 21320358]
6. Weng C, Batres C, Borda T, Weiskopf NG, Wilcox AB, Bigger JT, et al. A real-time screening alert improves patient recruitment efficiency. AMIA Annu Symp Proc 2011;2011:1489-1498 [FREE Full text] [Medline: 22195213]
7. Xu J, Rasmussen LV, Shaw PL, Jiang G, Kiefer RC, Mo H, et al. Review and evaluation of electronic health records-driven phenotype algorithm authoring tools for clinical and translational research. J Am Med Inform Assoc 2015 Nov;22(6):1251-1260 [FREE Full text] [doi: 10.1093/jamia/ocv070] [Medline: 26224336]
8. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015 Feb;53:162-173 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]
9. Sarmiento R, Dernoncourt F. Improving patient cohort identification using natural language processing. In: Secondary Analysis of Electronic Health Records. Cham (CH): Springer; Sep 10, 2016:405-417.
10. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
11. Hanauer DA, Barnholtz-Sloan JS, Beno MF, Del Fiol G, Durbin EB, Gologorskaya O, et al. Electronic Medical Record Search Engine (EMERSE): an information retrieval tool for supporting cancer research. JCO Clin Cancer Inform 2020 May;4:454-463 [FREE Full text] [doi: 10.1200/CCI.19.00134] [Medline: 32412846]
12. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: a report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform 2015 Jun;55:290-300 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.003] [Medline: 25979153]
13. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
14. Wagholikar KB, Ainsworth L, Vernekar VP, Pathak A, Glynn C, Zelle D, et al. Extending i2b2 into a framework for semantic abstraction of EHR to facilitate rapid development and portability of Health IT applications. AMIA Jt Summits Transl Sci Proc 2019;2019:370-378 [FREE Full text] [Medline: 31258990]
15. Tao S, Cui L, Wu X, Zhang G. Facilitating cohort discovery by enhancing ontology exploration, query management and query sharing for large clinical data repositories. AMIA Annu Symp Proc 2017;2017:1685-1694 [FREE Full text] [Medline: 29854239]
16. Garde S, Chen R, Leslie H, Beale T, McNicoll I, Heard S. Archetype-based knowledge management for semantic interoperability of electronic health records. Stud Health Technol Inform 2009;150:1007-1011. [Medline: 19745465]
17. Leslie H. OpenEHR archetype use and reuse within multilingual clinical data sets: case study. J Med Internet Res 2020 Nov 02;22(11):e23361 [FREE Full text] [doi: 10.2196/23361] [Medline: 33035176]
18. Helou SE, Kobayashi S, Yamamoto G, Kume N, Kondoh E, Hiragi S, et al. Graph databases for openEHR clinical repositories. IJCSE 2019;20(3):281. [doi: 10.1504/ijcse.2019.103955]
19. Wang L, Min L, Wang R, Lu X, Duan H. Archetype relational mapping: a practical openEHR persistence solution. BMC Med Inform Decis Mak 2015 Nov 05;15:88 [FREE Full text] [doi: 10.1186/s12911-015-0212-0] [Medline: 26541142]
20. Frade S, Freire SM, Sundvall E, Patriarca-Almeida JH, Cruz-Correia R. Survey of openEHR storage implementations. 2013 Presented at: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems; June 22, 2013; Porto, Portugal p. 303-307. [doi: 10.1109/cbms.2013.6627806]
21. SNOMED International. URL: http://www.snomed.org/ [accessed 2021-07-14]
22. LOINC. URL: https://loinc.org/ [accessed 2021-07-14]

23. eHealth & ICNP™. International Council of Nurses. URL: https://www.icn.ch/what-we-do/projects/ehealth-icnptm [accessed 2021-07-14]
24. International Statistical Classification of Diseases and Related Health Problems (ICD). URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2021-07-14]
25. RxNorm. NIH National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/rxnorm/index.html [accessed 2021-07-14]
26. Atalag K, Yang HY, Tempero E, Warren JR. Evaluation of software maintain ability with open EHR: a comparison of architectures. Int J Med Inform 2014 Nov;83(11):849-859. [doi: 10.1016/j.ijmedinf.2014.07.006] [Medline: 25153769]
27. Expression Language. OpenEHR. URL: https://specifications.openehr.org/releases/LANG/latest/expression_language.html [accessed 2021-07-14]
28. What is Elasticsearch? Elastic. URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/elasticsearch-intro.html [accessed 2021-07-14]
29. Chinese Clinical Trial Register (ChiCTR). URL: https://www.chictr.org.cn/enIndex.aspx [accessed 2021-07-14]
30. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. Summit Transl Bioinform 2010 Mar 01;2010:46-50 [FREE Full text] [Medline: 21347148]
31. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010 Jun;43(3):451-467 [FREE Full text] [doi: 10.1016/j.jbi.2009.12.004] [Medline: 20034594]
32. Allen JF. Maintaining knowledge about temporal intervals. Commun. ACM 1983 Nov;26(11):832-843. [doi: 10.1145/182.358434]
33. ANTLR (ANother Tool for Language Recognition). URL: https://www.antlr.org/ [accessed 2021-07-14]
34. Java Persistence API. URL: https://jakarta.ee/specifications/persistence/ [accessed 2021-07-14]
35. Hibernate Search. URL: https://docs.jboss.org/hibernate/stable/search/reference/en-US/html_single/ [accessed 2021-07-14]
36. openEHR java-libs. URL: https://github.com/openEHR/java-libs [accessed 2021-07-14]
37. Rinaldi E, Thun S. From openEHR to FHIR and OMOP data model for microbiology findings. Stud Health Technol Inform 2021 May 27;281:402-406. [doi: 10.3233/SHTI210189] [Medline: 34042774]
38. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Med Inform Assoc 2019 Nov 01;26(11):1163-1171 [FREE Full text] [doi: 10.1093/jamia/ocz163] [Medline: 31562516]
39. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc 2019 Nov 01;26(11):1247-1254 [FREE Full text] [doi: 10.1093/jamia/ocz149] [Medline: 31512729]
40. Xiong Y, Shi X, Chen S, Jiang D, Tang B, Wang X, et al. Cohort selection for clinical trials using hierarchical neural network. J Am Med Inform Assoc 2019 Nov 01;26(11):1203-1208 [FREE Full text] [doi: 10.1093/jamia/ocz099] [Medline: 31305921]
41. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc 2011 Dec;18 Suppl 1:i116-i124 [FREE Full text] [doi: 10.1136/amiajnl-2011-000321] [Medline: 21807647]
42. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. J Am Med Inform Assoc 2017 Nov 01;24(6):1062-1071 [FREE Full text] [doi: 10.1093/jamia/ocx019] [Medline: 28379377]
43. Tseo Y, Salkola MI, Mohamed A, Kumar A, Abnousi F. Information extraction of clinical trial eligibility criteria. arXiv preprint arXiv:2006.07296 2020 Jun 12:1-4.
44. Chen M, Du F, Lan G, Lobanov VS. Using pre-trained transformer deep learning models to identify named entities and syntactic relations for clinical protocol analysis. 2020 Presented at: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1); March 25, 2020; Palo Alto, CA p. 1-8.
45. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc 2019 Apr 01;26(4):294-305 [FREE Full text] [doi: 10.1093/jamia/ocy178] [Medline: 30753493]
46. Wang P, Shi T, Reddy CK. Text-to-SQL generation for question answering on electronic medical records. 2020 Presented at: Proceedings of The Web Conference 2020; April 20-24, 2020; Taipei, Taiwan p. 350-361. [doi: 10.1145/3366423.3380120]

Abbreviations

AM: archetype model
ANTLR: ANother Tool for Language Recognition
API: Application Programming Interface
CDR: clinical data repository
ChiCTR: Chinese Clinical Trial Registry
CRF: case report form
EHR: electronic health record

EL: Expression Language
ES: Elasticsearch
i2b2: Informatics for Integrating Biology & the Bedside
ICD-10: International Classification of Diseases, Tenth Revision
JPA: Java Persistence API
MMSE: Mini-Mental State Exam
MoCA: Montreal Cognitive Assessment
NLP: natural language processing
RM: reference model
SQL: Structured Query Language

Edited by G Eysenbach; submitted 29.08.21; peer-reviewed by L Min, N Deng; comments to author 20.09.21; revised version received 27.09.21; accepted 27.09.21; published 21.10.21.

Please cite as:
Li M, Cai H, Nan S, Li J, Lu X, Duan H
A Patient-Screening Tool for Clinical Research Based on Electronic Health Records Using OpenEHR: Development Study
JMIR Med Inform 2021;9(10):e33192
URL: https://medinform.jmir.org/2021/10/e33192
doi:10.2196/33192
PMID:34673526

Original Paper

Expressiveness of an International Semantic Standard for Wound Care: Mapping a Standardized Item Set for Leg Ulcers to the Systematized Nomenclature of Medicine–Clinical Terms

Jens Hüsers[1], BSc, MA; Mareike Przysucha[1], BSc, MSc; Moritz Esdar[1], BA, MA; Swen Malte John[2], PhD; Ursula Hertha Hübner[1], PhD

[1]University of Applied Sciences Osnabrück, Osnabrück, Germany

[2]Institute for Interdisciplinary Dermatological Prevention and Rehabilitation, University of Osnabrück, Osnabrück, Germany

Corresponding Author:
Ursula Hertha Hübner, PhD
University of Applied Sciences Osnabrück
Albrechtstr 30
Osnabrück, 49076
Germany
Phone: 49 5419692012
Email: u.huebner@hs-osnabrueck.de

Abstract

Background: Chronic health conditions are on the rise and are putting high economic pressure on health systems, as they require well-coordinated prevention and treatment. Among chronic conditions, chronic wounds such as cardiovascular leg ulcers have a high prevalence. Their treatment is highly interdisciplinary and regularly spans multiple care settings and organizations; this places particularly high demands on interoperable information exchange that can be achieved using international semantic standards, such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT).

Objective: This study aims to investigate the expressiveness of SNOMED CT in the domain of wound care, and thereby its clinical usefulness and the potential need for extensions.

Methods: A clinically consented and profession-independent wound care item set, the German National Consensus for the Documentation of Leg Wounds (NKDUC), was mapped onto the precoordinated concepts of the international reference terminology SNOMED CT. Before the mapping took place, the NKDUC was transformed into an information model that served to systematically identify relevant items. The mapping process was carried out in accordance with the ISO/TR 12300 formalism. As a result, the reliability, equivalence, and coverage rate were determined for all NKDUC items and sections.

Results: The developed information model revealed 268 items to be mapped. Conducted by 3 health care professionals, the mapping resulted in *moderate* reliability (κ=0.512). Regarding the two best equivalence categories (symmetrical equivalence of meaning), the coverage rate of SNOMED CT was 67.2% (180/268) overall and 64.3% (108/168) specifically for wounds. The sections *general medical condition* (55/66, 83%), *wound assessment* (18/24, 75%), and *wound status* (37/57, 65%), showed higher coverage rates compared with the sections therapy (45/73, 62%), wound diagnostics (8/14, 57%), and patient demographics (17/34, 50%).

Conclusions: The results yielded acceptable reliability values for the mapping procedure. The overall coverage rate shows that two-thirds of the items could be mapped symmetrically, which is a substantial portion of the source item set. Some wound care sections, such as *general medical conditions* and *wound assessment*, were covered better than other sections (*wound status*, *diagnostics*, and *therapy*). These deficiencies can be mitigated either by postcoordination or by the inclusion of new concepts in SNOMED CT. This study contributes to pushing interoperability in the domain of wound care, thereby responding to the high demand for information exchange in this field. Overall, this study adds another puzzle piece to the general knowledge about SNOMED CT in terms of its clinical usefulness and its need for further extensions.

(JMIR Med Inform 2021;9(10):e31980) doi:10.2196/31980

KEYWORDS

wound care; chronic wound; chronic leg ulcer; SNOMED CT; health information exchange; semantic interoperability; terminology mapping

Introduction

Background

Chronic health conditions are on the increase, constituting a long lasting disease burden for patients [1,2], and posing high economic pressure for health systems [3,4], as they require well-coordinated prevention and treatment. Among the chronic conditions, diabetes and vascular diseases causing chronic wounds are common [5]. The prevalence of chronic wounds is estimated to be 2.21 per 1000 people worldwide and is expected to increase [6]. There are different types of chronic wounds depending on the primary disease and the site, for example, diabetic foot ulcers, leg ulcers, and pressure ulcers. Among them, leg ulcers constitute the most common chronic wounds [6]. Also known as ulcus cruris or chronic leg wound, a leg ulcer is a skin defect located on the lower leg or the foot that fails to heal. Leg ulcers are caused by underlying cardiovascular diseases, most often peripheral arterial occlusive disease or venous insufficiency [7]. Furthermore, they are associated with severe complications, such as pain, immobility, local and systemic infections, and even amputations [8].

As with most chronic diseases, an interdisciplinary regimen promises to be an effective clinical therapy strategy for chronic leg wounds [9,10]. To achieve optimal treatment results for patients with wounds, physicians from multiple medical fields and disciplines such as dermatology, cardiology, surgery, primary care, specialized wound care nurses, and physical therapists are part of the interdisciplinary team [11]. In addition to the interprofessional nature of chronic wound care, it is also highly interorganizational, as patients with leg ulcers are most often treated in ambulatory settings with multiple health care providers involved, such as ambulatory nursing organizations, physician offices, and specialized wound care facilities [2].

The characteristics of chronic wound care, and its interdisciplinary and interorganizational aspects, demand information exchange about the patients between the attending medical professionals to coordinate and manage care. The healing time and disease burden can thereby be curtailed [12].

Wound Care and Information Interoperability

To standardize and improve the documentation and communication process, various data sets for wound care were defined. Among them are the minimum data set for generic wound assessment [13], the electronic wound summary [14], and the National Consensus for the Documentation of Leg Ulcer (NKDUC) [15,16]. The minimum data set is a proposed item set for generic wound assessment in England. The electronic wound summary is a German national standard under development that describes the structure and type of information exchange at patient discharge. The NKDUC item set was published by the Institute for Health Services Research Dermatology and Nursing at the University Medical Centre Hamburg-Eppendorf, Germany. A representative consortium of clinicians, health care delivery organizations, and health insurers developed and consented this data set to standardize the assessment and record-keeping of chronic leg wounds. The data set considers international literature and medical guidelines [15,16] and is also applicable outside of Germany. The NKDUC is a comprehensive data set that goes beyond biomedical information and aims to describe the patient holistically.

Designed by the relevant professions, consented data sets such as the NKDUC are the prerequisite for meaningful record-keeping and facilitate interorganizational and interprofessional information exchange, which is a crucial process in the domain of wound care. Health information technology (HIT) systems such as electronic health records, patient health records, and patient portals can unfold the full potential of clinically consented data sets as they enable real-time, location-independent information exchange that complies with patient data protection regulations [17]. However, information sharing across HIT systems with different data models requires a standardized semantic representation of the data set's content so that the systems can process the received clinical information appropriately [18,19]. A translation of the elements of a data set into a reference terminology guarantees semantic interoperability across HIT systems, thereby incentivizing adoption and data sharing.

Aiming for semantic interoperability in health care, the International Health Terminology Standards Development Organisation (IHTSDO) publishes and maintains the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT). SNOMED CT promises to be the leading reference terminology for clinical terms on a global scale that enables health care professionals to exchange the semantics of data, that is, its clinical meaning [20]. SNOMED CT is a multi-hierarchical terminology that represents a clinical idea as a single SNOMED CT concept with a unique SNOMED CT identifier. Compared with other terminologies such as International Disease Classification Version 10 (ICD-10), which focuses on disease classification, SNOMED CT aims to describe the complete domain of health care entities with a high degree of specificity interrelating the concepts [21]. Hence, it promises to have the potential to cover the information elements of wound care. Furthermore, besides communication and record-keeping, standardized and interoperable clinical documentation from multiple clinical sites constitute the backbone of data-driven clinical decision support systems that may further improve treatment outcomes [18]. In recent years, an increasing number of countries have adopted SNOMED CT at the national level, including Switzerland and Austria, which joined IHTSDO in 2016 and 2018, respectively. Germany also followed suit in 2021 [22].

SNOMED CT has proven its usefulness in different areas of health care, such as, trauma information in emergency medicine records [23], cancer registries [24], and cardiology [25]. SNOMED CT therefore lends itself to be tested for chronic

wounds as well to provide solutions to foster interdisciplinary leg wound care [26]. However, there is limited scientific research and evidence of how well chronic wounds can be coded in SNOMED CT; to our knowledge, there are only 2 studies. The first study mapped a set of 13 items relevant to pressure ulcers on SNOMED CT [27]. In the second study, a regional wound assessment document comprising 116 items was mapped using SNOMED CT, showing a coverage rate of 50.6% [28]. However, both studies were limited to the investigation of nursing-specific source documents. Therefore, mapping of a profession-independent, more comprehensive, and interdisciplinary consented data set to study the expressiveness of SNOMED CT promises to be rewarding in terms of revealing the potentials and limits of SNOMED CT. Such studies may also provide evidence and valuable insights for stakeholders on the use of SNOMED CT and for realizing interoperability in wound care. In addition, they might motivate clinicians to discover SNOMED CT and its usability in specific domains.

Objective and Research Questions

The principal objective of this study is to investigate the rate at which SNOMED CT—using precoordinated concepts—covers the medically relevant expressions and terms used in the care of people with chronic wounds, particularly with chronic leg wounds. This procedure should provide evidence on the expressiveness of SNOMED CT in this medical domain and, therefore, its clinical usefulness and the potential need for extensions. Accordingly, this study pursued the following research questions:

1. Which leg ulcer concepts should be matched with SNOMED CT?
2. What is the reliability of the mapping process?
3. What is the coverage rate of SNOMED CT for leg wound terms and expressions, that is, how many source items are present in SNOMED CT?

Methods

Wound Care Item Set and General Methodology

To test the expressiveness and clinical usefulness of SNOMED CT in the care of patients with chronic wounds and chronic leg wounds in particular, a wound care item set based on international medical guidelines and standards with a high degree of clinical acceptance is needed. Consequently, it was decided to use the NKDUC mentioned above, which is a good example of a clinical data set to serve pars pro toto for others. The decision was made on the grounds that it is a standardized data set drawing on international recommendations and thus, features the necessary validity [15,16]. Furthermore, it embraces a rich set of terms mirroring the wound assessment, wound status, diagnostic measures, and treatments.

To meet the research objective and answer the research questions, this study comprised three main consecutive methodological blocks: First, a formal information model based on the NKDUC was designed. Second, mapping was conducted according to the Technical Report 12300:2014 of the ISO [29]. Its 21 mapping principles guided the entire mapping process (Multimedia Appendix 1). Finally, the coverage rate was determined.

Information Model

The information model was developed using all NKDUC items of the following sections: *patient demographics*, *general medical condition*, *wound assessment*, *wound status*, *diagnostics*, and *therapy*. Therefore, only NKDUC items that were consented by the NKDUC consortium in a Delphi-based process were used [15]. Other sections were excluded, that is, *patient-related outcomes*, *patient education*, and *nutritional status*, as they exclusively referred to external sources, such as questionnaires, for example Wound-QoL (Questionnaire on quality of life with chronic wounds) [30].

Created mainly for a clinical audience, the NKDUC has a flat tabular structure. Classes and class attributes of the information model were derived from this structure, including value sets consented by the NKDUC consortium as enumerations. The information model used the class diagram notation of the Unified Modeling Language. On the basis of this model, all items (ie, class names, attributes, and value sets) constitute the set of items that were mapped to SNOMED CT.

Wherever possible, we aimed to reduce the redundant information of the NKDUC introduced through its hierarchical structure. For example, the NKDUC item *metabolic disorders* contains further detailed items such as *diabetes mellitus* and *hyperuricemia*. SNOMED CT already contains this relationship through its taxonomy, governed by the concept model [31]. In this case, we included both conditions, that is, *diabetes mellitus* and *hyperuricemia*, sparing their parent concept.

Mapping

In the second methodological block, the NKDUC items represented in the information model were mapped onto the target terminology SNOMED CT. Mapping of the items was performed as a nonautomatic procedure. As such, it was manually conducted by 3 clinicians, that is, nurses experienced in wound care with a master's degree in health management and a major in health informatics. Before the mapping process started, all 3 nurses were trained to work with the SNOMED CT. The training mainly focused on the logical model, which provides the fundamental structure of SNOMED CT, and the concept model, which specifies both the top-level concepts (ie, hierarchies) and the arrangement of concepts within and between these hierarchies. The mapping was conducted in 2019 using the international SNOMED CT version (January 2019) and the IHTSDO SNOMED CT browser. In this study, only precoordinated SNOMED CT concepts were used for the mapping to scrutinize the coverage rate.

Each of the 3 nurses mapped the complete NKDUC. First, they translated each source item into English. Each mapper was then instructed to use this translation to identify the semantic equivalent concept from the target terminology. As SNOMED CT concepts may share meaning but differ in granularity, the mappers were advised to scan the hierarchy of identified concepts and select a single concept that provides the highest semantic intersection between the source item and the target concept. As a result, each mapper created a simple reference

set: a one-to-one relationship between the NKDUC and SNOMED CT [29].

To answer the second research question, the interrater reliability of the mapping was assessed by computing the Fleiss κ statistic. This statistic quantifies the concordance between the maps, expressing the reliability with a number ranging within the closed interval from zero to one ($0 \leq \kappa \leq 1$), with high Fleiss κ values reflecting a high agreement between the mappers [32,33]. The advantage of this statistic is that it acknowledges that agreement occurs randomly and accounts for it so that its estimate is more robust than the proportional agreement [34]. This assessment was performed for the overall mapping and the six distinct sections of the NKDUC.

Equivalence Rating and Coverage Rate

The third methodological block included three successive steps: (1) the semantic equivalence of the previously created maps was rated, (2) the final concept was chosen, and (3) the coverage rate of the mapping was calculated (research question 3).

The equivalence rating was conducted according to the scheme described by ISO/TR 12300. According to this scheme, the semantic equivalence of the map, expressing its quality, was categorized using five degrees (Multimedia Appendix 2). The first degree describes the semantic equivalence of meaning (lexical as well as conceptual); the second degree does so too, but with synonymy. The third and fourth degrees indicate that both concepts share meaning; the former describes a broader meaning of the source concept, whereas the latter has a narrower meaning. Finally, the fifth degree indicates that mapping is impossible to achieve, as the target terminology lacks concepts that share meaning with the source concept.

The equivalence rating was independently conducted by two assessors (JH and MP), both experienced in health informatics and interoperability in wound care. Again, the reliability of the equivalence rating between both assessors was evaluated using the Fleiss κ statistic. Both assessors then selected the final concept and defined the final equivalence rating in a joint discussion. This procedure resulted in a final map for which the coverage rates were calculated.

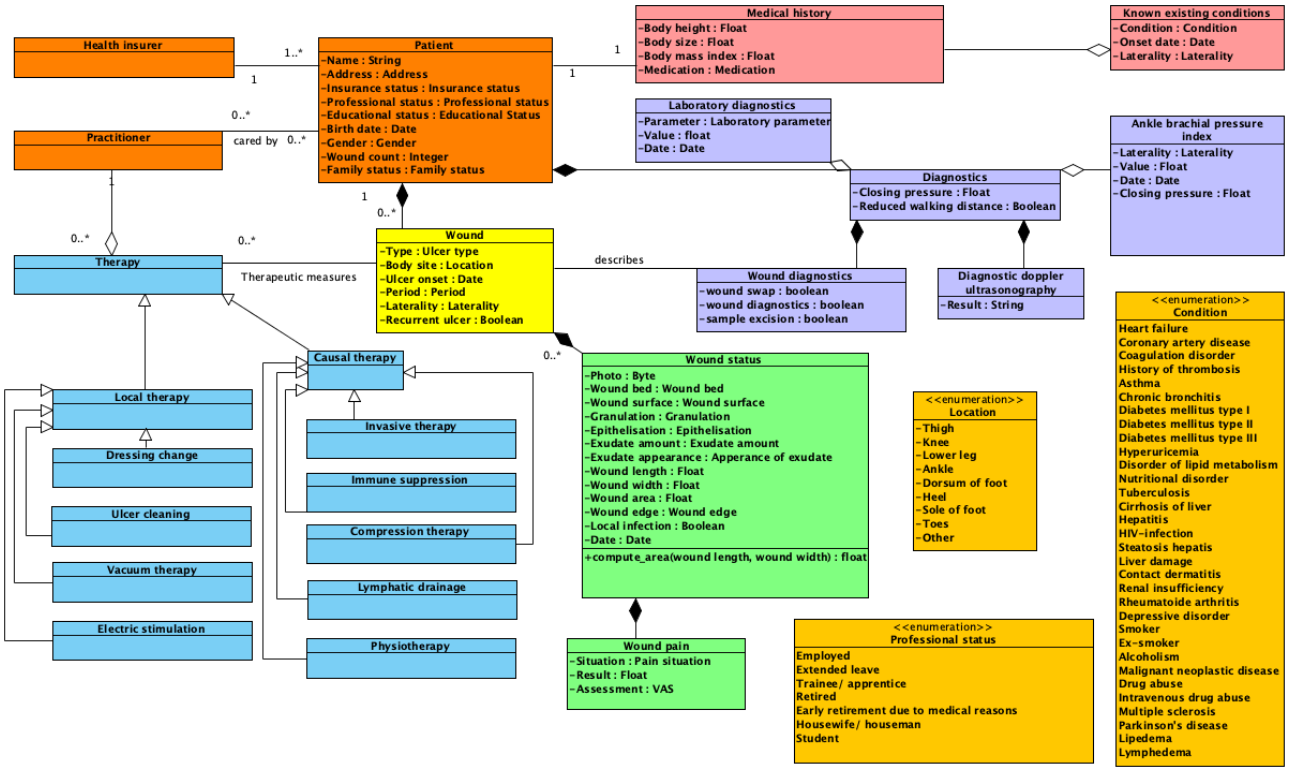
Data were entered and stored in Microsoft Excel. Data processing and analysis were performed using the Python programming language and additional open-source packages [35,36]. The data and Python scripts are available on the web (Multimedia Appendix 2).

Results

Information Model

The information model derived from the NKDUC revealed 268 distinct items for mapping. It included 25 classes, 66 attributes, 23 value sets, and 23 relations. Figure 1 shows an overview of the information model. The complete model is shown in Multimedia Appendix 2.

Figure 1. The information model (unified modelling language class diagram), with a subset of 3 selected value sets shown in dark-yellow (ulcer location, ulcer type, and condition). The class diagram only showcases the attributes for selected classes, for example, patient, wound, and wound status (orange: patient demographics, light red: general medical condition, purple: wound diagnostics, green: ulcer status, yellow: ulcer assessment, and sky-blue: therapy). VAS: visual analog scale.



Mapping and Reliability Rating

In the second methodological block, the previously identified 268 items were mapped first then the reliability was analyzed (Table 1). Regarding the overall reliability, the Fleiss κ value was κ=0.512. This value falls in the range of 0.41 and 0.60, which, according to Landis et al [37], can be considered a *moderate* degree of agreement. In addition, we analyzed the reliability of the mapping for each section of the NKDUC. In this context, the three mappers were most concordant for the section *general medical condition* with κ=0.754, which is considered a *substantial* agreement. The section *Patient demographics* and *Wound assessment* showed *moderate* agreement. The raters were most discordant for the sections *wound status* (κ=0.366), *therapy* (κ=0.367), and *diagnostics* (κ=0.280), which is considered as *fair* agreement [36].

Table 1. Interrater reliability of the three mappers that conducted the mapping (N=268).

	Reliability: Fleiss κ	Number of items, n (%)
Patient demographics	0.575	34 (12.7)
General medical condition	0.754	66 (24.6)
Wound assessment	0.568	24 (9)
Wound status	0.366	57 (21.3)
Diagnostics	0.280	14 (5.2)
Therapy	0.367	73 (27.2)
Overall	0.512	268 (100)

Equivalence Rating and Coverage Rate

In the third methodological block, the map was finalized on the basis of equivalence rating methodology. According to the Fleiss κ value of κ=0.702, the reliability of the equivalence rating of the 2 assessors can be considered a *substantial* agreement [36]. The reliability values of the equivalence rating per section are provided in Multimedia Appendix 3.

On the basis of the equivalence rating, 79.1% (212/268) of the NKDUC items had a match in the target terminology SNOMED CT (Table 2). With respect to distinct degrees of equivalence, 43.7% (117/268) NKDUC items shared lexical and conceptual meaning (degree 1) and 23.5% (63/268) items shared meaning through synonyms (degree 2), yielding a total of 67.2% (180/268) items in the two highest categories. Furthermore, 11.9% (32/268) NKDUC items matched to SNOMED CT with semantic asymmetry combining degrees 3 and 4, among which 2.2% (6/268) items had a broader (degree 3) and 9.7% (26/268) items had a narrower (degree 4) meaning than SNOMED CT. Finally, 20.9% (56/268) items remained unmatched, described as degree 5 in the ISO/TR 12300 equivalence scheme. Table 2 breaks down the coverage rates according to equivalence categories and Table 3 shows examples for each equivalence category (see Multimedia Appendix 4 for the detailed coverage rates per equivalence category).

Table 2. The mapping coverage rate is presented using ISO/TR 12300 equivalence categories for the complete National Consensus for the Documentation of Leg Wounds and its sections (N=268).

Equivalence categories	Overall (n=268), n (%)	Section, n (%)					
		01: Patient demographics (n=34)	02: General medical condition (n=66)	03: Wound assessment (n=24)	04: Wound status (n=57)	05: Diagnostics (n=14)	06: Therapy (n=73)
Semantic symmetric match present (degrees 1 and 2)	180 (67.2)	17 (50)	55 (83.3)	18 (75)	37 (64.9)	8 (57.1)	45 (61.6)
Semantic asymmetry match present (degrees 3 and 4)	32 (11.9)	5 (14.7)	4 (6.1)	1 (4.2)	7 (12.3)	3 (21.4)	12 (16.4)
Semantic match absent (degree 5)	56 (20.9)	12 (35.3)	7 (10.6)	5 (20.8)	13 (22.8)	3 (21.4)	16 (21.9)

Table 3. Examples of the equivalence rating.

Degree of equivalence (ISO/TR 12300)	Source (NKDUC[a])	Target	
		SNOMED CT[b] (descriptor)	(SCTID[c])
Equivalence of meaning; lexical, as well as conceptual	Ankle-brachial pressure index	Ankle-brachial pressure index (observable entity)	446841001
Equivalence of meaning, but with synonymy	Ulcus cruris arteriosum	Arteritic leg ulcer (disorder)	402862000
Source concept is broader and has a less specific meaning than the target concept	Skin condition	Periwound skin condition (observable entity)	700149001
Source concept is narrower and has a more specific meaning than the target concept	Wound infection	Infection status (observable entity)	405009004
No map is possible	Extent of wound area	N/A[d]	N/A

[a]NKDUC: National Consensus for the Documentation of Leg Wounds.

[b]SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms.

[c]SCTID: Systematized Nomenclature of Medicine–Clinical Terms Identifier.

[d]N/A: not applicable.

The overall coverage rate (covering degrees 1 and 2) was 67.2% (180/268). When considering the semantic equivalence for each section distinctly, the sections *general medical condition*, *wound assessment*, and *wound status* had the highest coverage rates, that is, 83% (55/66), 75% (18/24), and 65% (37/57), respectively. In contrast, the section *patient demographics* had the lowest coverage rate (17/34, 50%) and the highest number of unmatched items (12/34, 35%; Table 2). Regarding wound-specific information covered according to the two best degrees (1 and 2), *wound assessment* ranked first (18/24, 75%) and *wound diagnostics* last (8/14, 57%), with *wound status* and *therapy* ranking in between. The overall wound-specific coverage rate was 64.3% (108/168). When considering symmetric and asymmetric coverage (degrees 1-4), the wound-specific overall rate increased to 78% (131/168).

Discussion

Principal Findings

This study investigated the expressiveness of SNOMED CT in the domain of chronic wounds. It presents a mapping according to the ISO/TR 12300 formalism between the internationally grounded and nationally consented German data set NKDUC and the international terminology SNOMED CT. The NKDUC-based information model developed before the mapping revealed 268 items to be mapped. Conducted by 3 health care professionals, the mapping showed *moderate* reliability (κ=0.512). The coverage rate of SNOMED CT was 67.2% (180/268; symmetric match) overall and 64.3% (108/168) specifically for wounds.

Coverage Rate

The achieved coverage rate can be regarded as satisfactory as there is a direct, symmetric match in SNOMED CT for two-thirds of all mapped items (180/268, 67.2%, degrees 1 and 2). An additional 11.9% (32/268) of the mapped items received a rating of an asymmetric match (degrees 3 and 4), which adds to 79.1% (212/268) coverage. Wound assessment ranked first (18/24, 75%) and wound diagnostics ranked last (8/14, 57%).

The mapping of a regional, nurse-specific wound care data set from the United Kingdom, Columbia, and Canada yielded a coverage rate of 50.7% [28]. In comparison, the overall coverage rates on the basis of precoordinated concepts in other clinical domains, for example emergency medicine (89%) [19], were higher but could also be as low as 30% in the case of the human phenotype ontology [38]. In this context, the coverage rates in this mapping could be deemed satisfactory.

Considering the distinct NKDUC sections, heterogeneous coverage rates became apparent. At the end of both extremes, the section *general medical condition* had the highest coverage with over 83% (55/66) and the section *patient demographics* had the lowest coverage at 50% (17/34). The former section (*general medical condition*) mainly contains a list of ulcer-relevant conditions also found in the ICD-10 classification. Past ICD-10 mappings with SNOMED CT showed that it covers those items generally well [39], which explains the high coverage rate of this section. The latter section (*patient demographics*) contains German-specific items, such as educational, marital, professional, and health insurance status, for which few matching concepts were identified in SNOMED CT. The reduced coverage rate in this section reflects the fact that the mapping was performed using the international SNOMED CT version for German-specific items. Our findings support the need to fill these gaps in the German-specific items for a national German SNOMED CT version.

Although the wound-specific sections (ie, *wound status*, *wound assessment*, *diagnostics*, and *therapy*) showed a fair to reasonable coverage rate, the mappings thereof remained incomplete. For example, 57% (8/14) of the items in the *diagnostics* sections could be mapped literally to the same term or synonym.

In addition, our investigation revealed gaps in expressing wound-care-specific terms in SNOMED CT. Clinically, interdisciplinary data sets that are based on the literature and consented by medical experts, such as the NKDUC, provide valuable insights for identifying and filling these gaps.

One approach to this is post coordination, which is used across different domains such as cardiology and clinical phenotyping data [18]. Governed by SNOMED CT's compositional grammar and by composing existing SNOMED CT concepts, postcoordination generates semantically equivalent expressions for source terms that are unavailable as precoordinated concepts in SNOMED CT. Therefore, SNOMED CT expands its semantics and can fill gaps in a map. Postcoordination seems especially promising when the target concepts have a broader meaning (degree 4), as postcoordination can narrow down the meaning of existing concepts. Moreover, even for missing matches (degree 5), postcoordination offers a solution. For example, postcoordination would lead to the following SNOMED CT expression to code a *leg ulcer smear procedure*:

> *(16314007 |Microbial smear examination (procedure)|: {363700003 |Direct morphology (attribute)| = 56208002 |Ulcer (morphological abnormality)|, 363704007 |Procedure site (attribute)| = 416077002 | skin and/or subcutaneous tissue structure of lower limb (body structure)|}).*

However, postcoordination is more complex and requires more effort from implementers [40], and using precoordinated concepts may facilitate the implementation. Hence, maps to precoordinated concepts promise faster adoption than their postcoordinated counterparts. Despite these disadvantages, the findings of this study reveal the need for postcoordination. Future initiatives are necessary to cover the entire domain of wound care in SNOMED CT. These initiatives require a rigorous consensus-building process to generate and validate the concepts for clinical use. In this study, our findings suggest that the wound-specific sections *diagnostics* and *wound status* may benefit the most from postcoordination as they showed the lowest coverage rates. However, when postcoordination fails, missing concepts should be added to SNOMED CT, for example, German-specific items mentioned above. Furthermore, new concepts to describe the progress of epithelization and granulation status to record wound healing are good examples to illustrate this need.

Both approaches, postcoordination and adding missing concepts, promise to close the semantic gaps identified in SNOMED CT and would allow NKDUC, and probably other documentation standards in wound care, to reach semantic interoperability. In summary, the findings show that SNOMED CT is partially ready for use in wound care documentation. However, further measures, such as those explained above, seem desirable.

Reliability

The strength of these findings depends highly on the reliability of mapping. The overall reliability of κ=0.512 is what the reference literature describes as a *moderate* agreement between the mappers. This statistic indicates that the findings of the mapping stand on solid ground.

However, in this mapping process, rather than selecting items from a small set of options, the raters had to choose from a vast range of SNOMED CT concepts, as it provides over 350,000 precoordinated concepts. This circumstance makes it generally more difficult to find a consensus, especially for source items where similar target concepts are available. This conclusion is supported by the fact that the Fleiss κ statistic tends to decrease as the number of categories increases [41]. To increase reliability, intermediary discussions among mappers would have been beneficial. However, all mappers followed the same mapping rules to support the reliability of the mapping.

When comparing different NKDUC sections, heterogeneous reliability values could be found. For example, mappers were more discordant for concepts concerning *diagnostics* compared with those concerning *general medical conditions*. This situation may imply that sections showing low reliability are challenging to map, either because there are many similar SNOMED CT concepts with high semantic overlap, the NKDUC items are ambiguous, or both cases hold true. Reliability values and coverage rates seem to be related as lower Fleiss κ values for *wound status* (0.366), *diagnostics* (0.280), and *therapy* (0.367) tend to correspond with lower coverage rates of 65% (37/57), 57% (8/14), and 62% (45/73), respectively. Similarly, higher reliability values for *general medical condition* (0.754) and *wound assessment* (0.568) vary with higher coverage rates of 83% (55/66) and 75% (18/24), respectively. Therefore, in either case, the mapped SNOMED CT concepts in the sections with lower reliability must be validated carefully before their use in clinical practice.

Information Model

Although the information model derived from the NKDUC primarily served as a source for identifying the items to be mapped, it also allows statements about the general validity of NKDUC by comparing this information model with others. For example, the openEHR template *wound assessment panel* and *wound presence assertions* [42], which partly represent wound phenomena, embrace similar content as the corresponding parts of the NKDUC information model. This overlap hints at the validity of this information model as well as the mapping and implications for SNOMED CT. Furthermore, it seems promising to integrate the identified SNOMED CT concepts into wound-specific openEHR archetypes and templates to enhance interoperability in the domain of chronic wound care [43].

Limitations

There are some limitations to be considered when interpreting the results. Most importantly, as mentioned above, this study did not make use of postcoordination, which most likely limited a higher coverage rate as postcoordination usually extends the content of SNOMED CT through compositional expressions [44]. However, this study was conducted to investigate the predefined content and its coverage rate in the ulcer care domain using the NKDUC as an example of a national consented collection of ulcer-relevant items. We plan to implement postcoordination for an upcoming mapping of the NKDUC to SNOMED CT to further fill semantic gaps and improve the coverage rate required for actual implementation in systems used in clinical care. Furthermore, as NKDUC focuses on cardiovascular leg wounds, the coverage rate for items of further wound types, such as pressure ulcers, must be investigated additionally.

Another limitation of this study is the absence of a German SNOMED CT version, which necessitates an intermediary translation by each mapper, which may have introduced bias. To avoid this bias and increase interrater reliability, an additional validation step between the translation of the source concept and its mapping to SNOMED CT may have been beneficial. However, differences in the translations tend to be less biasing when synonyms are available, as is the case for SNOMED CT, and can cover these different translations. To overcome the lack of a German SNOMED CT version, a German translation group was formed by the Swiss, Austrian, and German National Release Centers in 2021 to develop guidelines and initiate translation projects [45]. The mapping that was developed in this study may support the German translation group and contribute to the translation for the domain of wound care.

Conclusions

This study investigates the expression of SNOMED CT in the wound care domain based on a comprehensive, clinically consented data set. The results encourage the use of SNOMED CT and build the foundation for semantically interoperable systems to foster information exchange, which is crucial in the interprofessional and interorganizational setting of chronic wound care. Furthermore, the mapping followed the instructions of ISO/TR 12300 and determined the reliability as well as the equivalence and coverage rate. We thereby showcase a replicable procedure that can be used as a blueprint to produce reliable and transparent mappings in other domains as well.

Overall, this study adds another puzzle piece to the general knowledge about SNOMED CT in terms of its clinical usefulness and its need for further extensions. Semantic interoperability through SNOMED CT has become the most powerful in interdisciplinary and interprofessional scenarios across care settings, of which wound care is an excellent example.

Acknowledgments

This study was funded by the German Federal Ministry of Education and Research (grant: 13GW0171B). The funders had no role in the study design, data collection, analysis, decision to publish, or manuscript preparation. This study is the result of a close collaboration between all the authors. UHH and JH were the first to initiate the project. Subsequently, all authors planned the study and general concept. MP and JH conducted equivalence ratings. All the authors were involved in the interpretation and discussion of the results. The authors would like to thank Anne Paul, MA; Philip Matysek, MA; and Lisa Nolte, MA for conducting the Systematized Nomenclature of Medicine–Clinical Terms mapping to the National Consensus for the Documentation of Leg Wounds. Furthermore, the authors wish to thank their colleagues of the German Federal Ministry of Education and Research Project PosiThera (grant: 13GW0171B). In particular, they would like to acknowledge the support of Björn Sellemann, Stefan Vogel, Stefanie Wache, Jendrik Richter, Karin Güttler, and Sebastian Zebbities.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The 21 mapping principles proposed by the International Organization for Standardization in its Technical Report (ISO/TR 12300) and their application in this study.
[DOCX File , 17 KB - medinform_v9i10e31980_app1.docx]

Multimedia Appendix 2
The information model, reference map and, to support reproducible research, Python script of the analysis.
[DOCX File , 12 KB - medinform_v9i10e31980_app2.docx]

Multimedia Appendix 3
Reliability of the equivalence assessment.
[DOCX File , 13 KB - medinform_v9i10e31980_app3.docx]

Multimedia Appendix 4
The coverage rate of the mapping for each equivalence category of the ISO/TR 12300 standard for each degree separately.
[DOCX File , 15 KB - medinform_v9i10e31980_app4.docx]

References

1. González-Consuegra RV, Verdú J. Quality of life in people with venous leg ulcers: an integrative review. J Adv Nurs 2011 May;67(5):926-944. [doi: 10.1111/j.1365-2648.2010.05568.x] [Medline: 21241355]

2. Herberger K, Rustenbach SJ, Haartje O, Blome C, Franzke N, Schäfer I, et al. Quality of life and satisfaction of patients with leg ulcers--results of a community-based study. Vasa 2011 Mar;40(2):131-138. [doi: 10.1024/0301-1526/a000083] [Medline: 21500178]
3. Nussbaum SR, Carter MJ, Fife CE, DaVanzo J, Haught R, Nusgart M, et al. An economic evaluation of the impact, cost, and medicare policy implications of chronic nonhealing wounds. Value Health 2018 Jan;21(1):27-32 [FREE Full text] [doi: 10.1016/j.jval.2017.07.007] [Medline: 29304937]
4. Barnsbee L, Cheng Q, Tulleners R, Lee X, Brain D, Pacella R. Measuring costs and quality of life for venous leg ulcers. Int Wound J 2019 Feb;16(1):112-121 [FREE Full text] [doi: 10.1111/iwj.13000] [Medline: 30289621]
5. Raghupathi W, Raghupathi V. An empirical study of chronic diseases in the united states: a visual analytics approach. Int J Environ Res Public Health 2018 Mar 01;15(3):431 [FREE Full text] [doi: 10.3390/ijerph15030431] [Medline: 29494555]
6. Martinengo L, Olsson M, Bajpai R, Soljak M, Upton Z, Schmidtchen A, et al. Prevalence of chronic wounds in the general population: systematic review and meta-analysis of observational studies. Ann Epidemiol 2019 Jan;29:8-15. [doi: 10.1016/j.annepidem.2018.10.005] [Medline: 30497932]
7. Dissemond J. [Chronic leg ulcers]. Hautarzt 2017 Aug;68(8):614-620. [doi: 10.1007/s00105-017-4010-8] [Medline: 28638953]
8. Bui UT, Edwards H, Finlayson K. Identifying risk factors associated with infection in patients with chronic leg ulcers. Int Wound J 2018 Apr;15(2):283-290 [FREE Full text] [doi: 10.1111/iwj.12867] [Medline: 29250935]
9. Lokaltherapie chronischer Wunden bei Patienten mit den Risiken periphere arterielle Verschlusskrankheit, Diabetes mellitus, chronisch venöse Insuffizienz. AWMF online. URL: https://www.awmf.org/leitlinien/detail/ll/091-001.html [accessed 2021-09-03]
10. Bakker K, Apelqvist J, Lipsky B, Van Netten JJ, International Working Group on the Diabetic Foot. The 2015 IWGDF guidance documents on prevention and management of foot problems in diabetes: development of an evidence-based global consensus. Diabetes Metab Res Rev 2016 Jan;32 Suppl 1:2-6. [doi: 10.1002/dmrr.2694] [Medline: 26409930]
11. Alavi A, Sibbald RG, Phillips TJ, Miller OF, Margolis DJ, Marston W, et al. What's new: management of venous leg ulcers: approach to venous leg ulcers. J Am Acad Dermatol 2016 Apr;74(4):627-40; quiz 641. [doi: 10.1016/j.jaad.2014.10.048] [Medline: 26979354]
12. Gupta S, Andersen C, Black J, de Leon J, Fife C, Lantis Ii JC, et al. Management of chronic wounds: diagnosis, preparation, treatment, and follow-up. Wounds 2017 Sep;29(9):S19-S36 [FREE Full text] [Medline: 28862980]
13. Coleman S, Nelson EA, Vowden P, Vowden K, Adderley U, Sunderland L, Improving Wound Care Project Board, as part of NHS England's Leading Change Adding Value Framework. Development of a generic wound care assessment minimum data set. J Tissue Viability 2017 Nov;26(4):226-240 [FREE Full text] [doi: 10.1016/j.jtv.2017.09.007] [Medline: 29030056]
14. Hübner U, Schulte G, Flemming D. Der elektronische Wundbericht als Grundlage für eine interprofessionelle Kommunikation in der intersektoralen. ResearchGate. 2016. URL: https://www.hs-osnabrueck.de/fileadmin/HSOS/Homepages/KeGL/Artikel_Der_elektronische_Wundbericht_als_Grundlage_fuer_eine_interprofessionelle_Kommunikation_in_der_intersektoralen_Versor.pdf [accessed 2021-09-03]
15. Heyer K, Herberger K, Protz K, Mayer A, Dissemond J, Debus S, Konsensusgruppe. [German national consensus on wound documentation of leg ulcer : part 1: routine care - standard dataset and minimum dataset]. Hautarzt 2017 Sep;68(9):740-745. [doi: 10.1007/s00105-017-4011-7] [Medline: 28681135]
16. Herberger K, Heyer K, Protz K, Mayer A, Dissemond J, Debus S, Konsensusgruppe. [German national consensus on wound documentation of leg ulcer : part 2: routine care - classification of variable characteristics]. Hautarzt 2017 Nov;68(11):896-911. [doi: 10.1007/s00105-017-4012-6] [Medline: 28681136]
17. Benson T, Grieve G. SNOMED CT. In: Principles of Health Interoperability. Health Information Technology Standards. Cham: Springer International Publishing; 2016.
18. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014 Feb;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
19. Duarte J, Castro S, Santos M, Abelha A, Machado J. Improving quality of electronic health records with SNOMED. Procedia Technol 2014;16:1342-1350. [doi: 10.1016/j.protcy.2014.10.151]
20. Moreno-Conde A, Moner D, Cruz WD, Santos MR, Maldonado JA, Robles M, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc 2015 Jul;22(4):925-934. [doi: 10.1093/jamia/ocv008] [Medline: 25796595]
21. González Bernaldo de Quirós F, Otero C, Luna D. Terminology services: standard terminologies to control health vocabulary. Yearb Med Inform 2018 Aug;27(1):227-233 [FREE Full text] [doi: 10.1055/s-0038-1641200] [Medline: 29681027]
22. Patientendaten-schutz-gesetz ermöglicht digitale Lösungen für Patienten. Deutscher Bundestag. URL: https://www.bundestag.de/dokumente/textarchiv/2020/kw27-de-patientendatenschutz-701808 [accessed 2021-07-25]
23. Brammen D, Dewenter H, Heitmann KU, Thiemann V, Majeed RW, Walcher F, et al. Mapping equivalence of German emergency department medical record concepts with SNOMED CT after implementation with HL7 CDA. Stud Health Technol Inform 2017;243:175-179. [Medline: 28883195]

24. Sanz X, Pareja L, Rius A, Gálvez J, Escribà JM, Esteban L, et al. How cancer registries can detect neoplasms in pathology laboratories that code with SNOMED CT terminology? An actual, simple and flexible solution. Int J Med Inform 2020 Sep;141:104167. [doi: 10.1016/j.ijmedinf.2020.104167] [Medline: 32554239]
25. Soguero-Ruiz C, Mora-Jiménez I, Ramos-López J, Quintanilla Fernández T, García-García A, Díez-Mazuela D, et al. An interoperable system toward cardiac risk stratification from ECG monitoring. Int J Environ Res Public Health 2018 Mar 01;15(3):428 [FREE Full text] [doi: 10.3390/ijerph15030428] [Medline: 29494497]
26. Edwards A, Hollin I, Barry J, Kachnowski S. Barriers to cross--institutional health information exchange: a literature review. J Healthc Inf Manag 2010;24(3):22-34. [Medline: 20677469]
27. Kim H, Park H. Development and evaluation of data entry templates based on the entity-attribute-value model for clinical decision support of pressure ulcer wound management. Int J Med Inform 2012 Jul;81(7):485-492. [doi: 10.1016/j.ijmedinf.2011.10.008] [Medline: 22079242]
28. Block L, Handfield S. Mapping wound assessment data elements in SNOMED CT. Stud Health Technol Inform 2016;225:1078-1079. [Medline: 27332492]
29. ISO/TR 12300:2014 Health informatics — Principles of mapping between terminological systems. ISO. 2014. URL: https://www.iso.org/standard/51344.html [accessed 2021-09-03]
30. Augustin M, Baade K, Herberger K, Protz K, Goepel L, Wild T, et al. Use of the WoundQoL instrument in routine practice: feasibility, validity and development of an implementation tool. Wound Med 2014 Jun 1;5:4-8 [FREE Full text] [doi: 10.1016/j.wndm.2014.04.001]
31. Concept model overview. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCEG/Concept+Model+Overview [accessed 2021-07-07]
32. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76(5):378-382. [doi: 10.1037/h0031619]
33. Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Med Res Methodol 2016 Aug 05;16:93 [FREE Full text] [doi: 10.1186/s12874-016-0200-9] [Medline: 27495131]
34. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
35. Van Rossum G, Drake F. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
36. pandas-dev/pandas: Pandas 1.3.2. Zendo. 2021. URL: https://zenodo.org/record/5203279#.YTH4VI4zbIU [accessed 2021-09-03]
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biom 1977 Mar;33(1):159. [doi: 10.2307/2529310]
38. Winnenburg R, Bodenreider O. Coverage of phenotypes in standard terminologies. In: Proceedings of the Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day". 2014 Presented at: Joint Bio-Ontologies and BioLINK ISMB'2014 SIG session "Phenotype Day"; Jul 11-12, 2014; Boston USA p. 41-44 URL: https://lhncbc.nlm.nih.gov/LHC-publications/pubs/Coverageofphenotypesinstandardterminologies.html
39. SNOMED CT to ICD-10-CM Map. National Library of Medicine. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html [accessed 2021-07-07]
40. Elkin PL, Brown SH, Lincoln MJ, Hogarth M, Rector A. A formal representation for messages containing compositional expressions. Int J Med Inform 2003 Sep;71(2-3):89-102. [doi: 10.1016/s1386-5056(03)00087-x]
41. Sim J, Wright C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005 Mar;85(3):257-268. [Medline: 15733050]
42. Incubator: FHIR/OpenEHR Archetype Development. Open EHR. URL: https://ckm.openehr.org/ckm/incubators/1013.30.9 [accessed 2021-07-07]
43. Harris M, Langford L, Miller H, Hook M, Dykes P, Matney S. Harmonizing and extending standards from a domain-specific and bottom-up approach: an example from development through use in clinical applications. J Am Med Inform Assoc 2015 May;22(3):545-552. [doi: 10.1093/jamia/ocu020] [Medline: 25670750]
44. SNOMED CT Expressions - SNOMED CT Starter Guide. SNOMED International. URL: https://confluence.ihtsdotools.org/display/DOCSTART/+SNOMED+CT+Expressions [accessed 2020-11-03]
45. German translation group (GTG). SNOMED International. URL: https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=113417116 [accessed 2021-08-12]

Abbreviations

HIT: health information technology
ICD-10: International Disease Classification Version 10
IHTSDO: International Health Terminology Standards Development Organisation
NKDUC: National Consensus for the Documentation of Leg Wounds
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms
Wound-QoL: Questionnaire on quality of life with chronic wounds

Edited by G Eysenbach; submitted 16.07.21; peer-reviewed by C Vorisek, T Macieira; comments to author 09.08.21; revised version received 16.08.21; accepted 24.08.21; published 06.10.21.

Please cite as:
Hüsers J, Przysucha M, Esdar M, John SM, Hübner UH
Expressiveness of an International Semantic Standard for Wound Care: Mapping a Standardized Item Set for Leg Ulcers to the Systematized Nomenclature of Medicine–Clinical Terms
JMIR Med Inform 2021;9(10):e31980
URL: https://medinform.jmir.org/2021/10/e31980
doi:10.2196/31980
PMID:34428171

Original Paper

Common Data Elements for Meaningful Stroke Documentation in Routine Care and Clinical Research: Retrospective Data Analysis

Sarah Berenspöhler[1], MD; Jens Minnerup[2], MD; Martin Dugas[3], MD, MSc; Julian Varghese[1], MD, MSc

[1]Institute of Medical Informatics, Westfälische Wilhelms-University Münster, Münster, Germany

[2]Department of Neurology with Institute of Translational Neurology, University Hospital Münster, Münster, Germany

[3]Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

Corresponding Author:
Sarah Berenspöhler, MD
Institute of Medical Informatics
Westfälische Wilhelms-University Münster
Albert Schweitzer Campus 1, Building A11
Münster, 48149
Germany
Phone: 49 251 83 55262
Fax: 49 251 83 52259
Email: s.berenspoehler@t-online.de

Abstract

Background: Medical information management for stroke patients is currently a very time-consuming endeavor. There are clear guidelines and procedures to treat patients having acute stroke, but it is not known how well these established practices are reflected in patient documentation.

Objective: This study compares a variety of documentation processes regarding stroke. The main objective of this work is to provide an overview of the most commonly occurring medical concepts in stroke documentation and identify overlaps between different documentation contexts to allow for the definition of a core data set that could be used in potential data interfaces.

Methods: Medical source documentation forms from different documentation contexts, including hospitals, clinical trials, registries, and international standards, regarding stroke treatment followed by rehabilitation were digitized in the operational data model. Each source data element was semantically annotated using the Unified Medical Language System. The concept codes were analyzed for semantic overlaps. A concept was considered common if it appeared in at least two documentation contexts. The resulting common concepts were extended with implementation details, including data types and permissible values based on frequent patterns of source data elements, using an established expert-based and semiautomatic approach.

Results: In total, 3287 data elements were identified, and 1051 of these emerged as unique medical concepts. The 100 most frequent medical concepts cover 9.51% (100/1051) of all concept occurrences in stroke documentation, and the 50 most frequent concepts cover 4.75% (50/1051). A list of common data elements was implemented in different standardized machine-readable formats on a public metadata repository for interoperable reuse.

Conclusions: Standardization of medical documentation is a prerequisite for data exchange as well as the transferability and reuse of data. In the long run, standardization would save time and money and extend the capabilities for which such data could be used. In the context of this work, a lack of standardization was observed regarding current information management. Free-form text fields and intricate questions complicate automated data access and transfer between institutions. This work also revealed the potential of a unified documentation process as a core data set of the 50 most frequent common data elements, accounting for 34% of the documentation in medical information management. Such a data set offers a starting point for standardized and interoperable data collection in routine care, quality management, and clinical research.

(JMIR Med Inform 2021;9(10):e27396) doi:10.2196/27396

KEYWORDS

common data elements; stroke; documentation

Introduction

Background

Stroke is the second most common cause of death worldwide and is the most important cause of permanent disability in adults [1]. Owing to the increasing aging of the population, a steady incidence rate would probably lead to an increasing number of people being affected by stroke in Germany in the next decades [1].

In addition, the treatment of the disease and the consequent damage generate immense costs for the health system and thus the population [2]. The high incidence and prevalence rate of this disease in Germany induces numerous studies, the creation of therapy guidelines and regulations for proper initial treatment of stroke patients, and preventive measures. Currently, there are more than 300 certified stroke units in Germany [3]. These stroke units have to comply to a multitude of certification criteria and pass regular audits [3]. Primary as well as standardized secondary prevention (including early recurrence prevention) are also of great importance, and procedures for treating carotid artery stenosis are regulated by the S3 guidelines [4].

What about information management regarding stroke patients are there sufficient standards and guidelines for hospitals as well? It is common knowledge that patient documentation is a time-consuming endeavor. The documentation already starts in the emergency room when basic patient data are collected, and the medical history and initial examination results are recorded. Subsequent examination results, vital parameters, and health changes were documented according to a fixed time schedule. Even after hospitalization, a large amount of data are recorded during follow-up examinations, rehabilitation, or clinical research. Medical documentation accounts for 25% of the physician's workload and takes up as much time as direct patient care [5].

Objective

Although there are clear certification requirements for stroke units, guidelines for therapy, extensive rehabilitation networks, secondary prevention, and numerous research papers, the question arises as to what measures are taken to improve the collection and processing of data. Standardization allows for a sound way of transferring data between departments and institutions, which saves time and money [6,7]. Do such standardizations or maybe even a transferable core data set already exist? Queried institutions denied questions regarding standardizations for routine clinical documentation, software, transmission interfaces, or a core data set; there is also no superordinate institution for coordinating the exchange of data. The certification criteria for stroke units specify certain examinations, examination scores, and therapy cycles; documentation is expected, but there are no specific requirements or standards.

In some federal states of Germany, the only clear requirement specific to information management concerning stroke units is mandatory participation in the stroke registry for quality assurance. Clinics must provide a form with 77 data elements for each patient and are required to complete at least 90% of these [8]. The data elements are revised annually and partially adjusted.

The association of German stroke registries, the Arbeitsgemeinschaft Deutschsprachiger Schlaganfall-Register (ADSR), is a voluntary association of regional quality assurance projects regarding stroke treatment. The ADSR was founded in 1999 with the objective of developing a standardized data collection for stroke cases. It creates regional as well as supraregional comparisons based on scientific, quality-related, and epidemiological viewpoints. There are yearly meetings for members to reconcile documentation forms and discuss uniform quality indicators. These quality indicators are developed by a multidisciplinary work group, including representatives of the Deutsche Schlaganfall Gesellschaft (German Stroke Society), the Deutsche Gesellschaft für Neurologie (German Association of Neurology) and the Stiftung Deutsche Schlaganfall-Hilfe (German Stroke Foundation). Approximately 300,000 data records are evaluated by the ADSR each year [9]. The documentation for the stroke registry is usually conducted with additional software, so data from routine documentation are not taken over. Redundant documentation performed in parallel for different applications increases the effort and susceptibility to errors.

There are also registries for quality assurance in the domain of early rehabilitation and rehabilitation in general. The Hessische Krankenhausgesellschaft (Hesse Hospital Association) and health insurance associations in Hesse have come to a contractual agreement in this regard, but supraregionally there is no obligation to participate.

Problems with existing documentation procedures are not new. For many years, medical information management has been analyzed and discussed in the field of health informatics. Thus, studies similar to this one exist for other diseases, such as for acute myeloid leukemia [10] or acute coronary syndrome [11]. Since 2015, the German Research Foundation has funded several projects that aim to establish an information infrastructure for research data. In this context, more than 500,000 additional data models were processed at the Institute of Medical Informatics in Münster [12].

So far, endeavors regarding information management for stroke patients in Germany have been covered, but what is the international situation with regard to this? Are there standards, interfaces for different medical documentation sources, or even a specification of a core data set? The National Institute of Neurological Disorders and Stroke (NINDS) is a research institute belonging to the National Institutes of Health and is supported by the Department of Health and Human Services. One of the objectives of the NINDS is to develop data standards for clinical research and accessible tools while improving data quality and cost control [13,14]. Data elements from case report forms, clinical routine forms, guidelines, and clinical data standards are available to identify core data elements [13,15-17].

There are also registries for patients who had stroke in other countries. Thus, data structures from the Austrian Stroke Registry were included in this study [18].

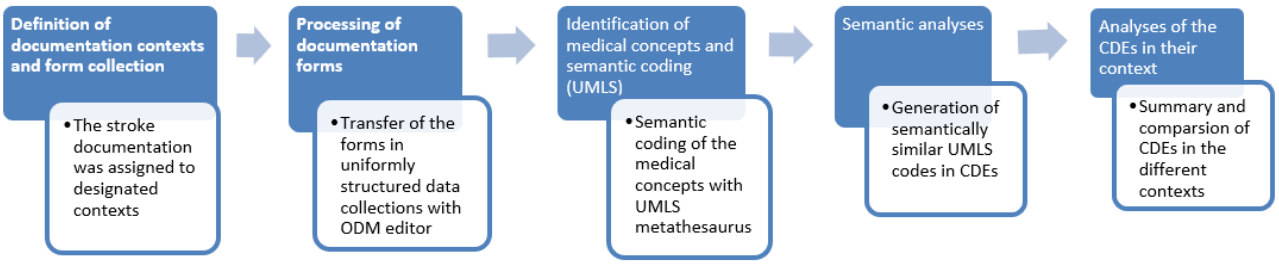
In summary, the objective of this work is to contribute to a crossdomain core data set for patients who had stroke that could function as a standard and be used for data exchange.

Methods

Overview of the Methods

Figure 1 illustrates the conducted main steps from the definition of documentation contexts over transfer of uniformly structured data and semantic coding to create and comparison of common data elements' (CDEs).

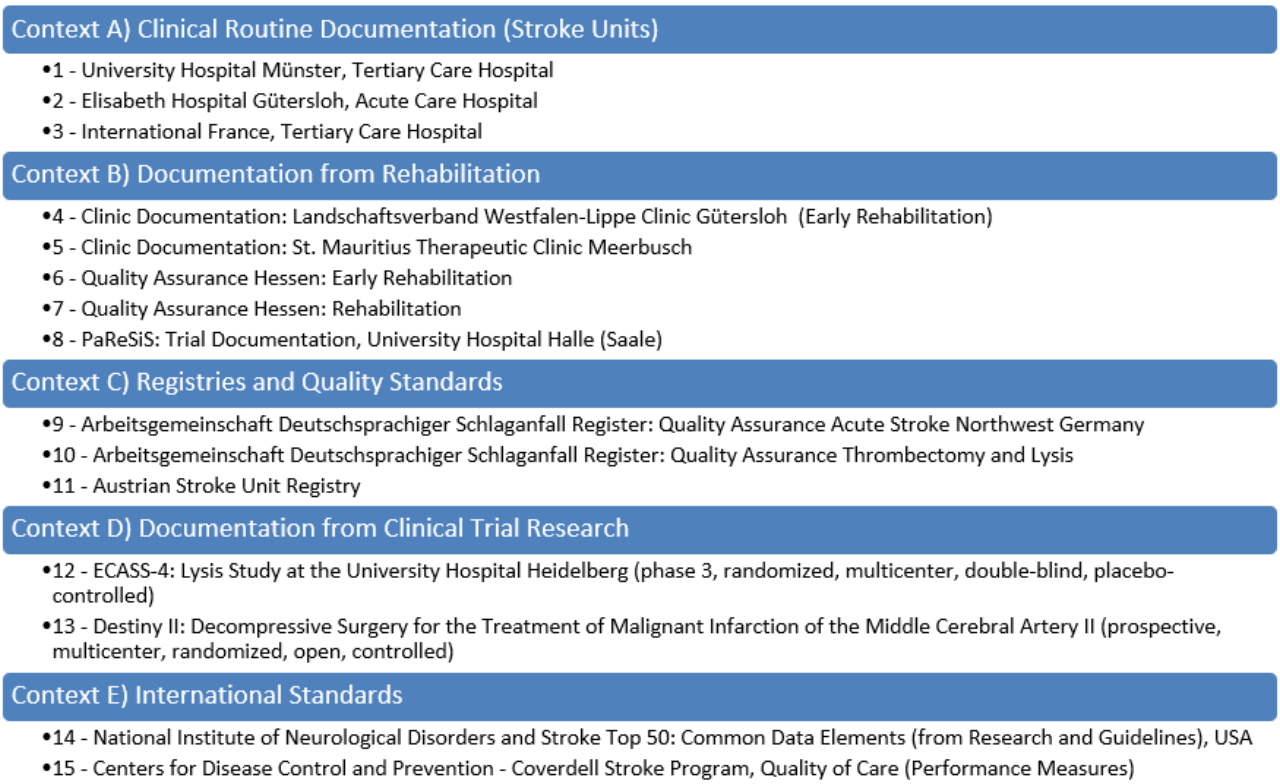
Figure 1. Overall workflow for definition of CDEs from existing documentation forms. CDE: common data element; ODM: operational data monitor; UMLS: Unified Medical Language System.



Definition of Documentation Contexts and Form Collection

Patient documentation forms for stroke patients were collected from hospitals, rehabilitation facilities, research papers, registries, and standards for the period from 2014 to 2017 (Figure 2). The selection was conducted by 2 medical experts including 1 clinical stroke expert based on the availability of documentation forms and broad content coverage in clinical research and care.

Figure 2. Designated contexts. ADSR: Arbeitsgemeinschaft Deutschsprachiger Schlaganfall-Register; CDC: Centers for Disease Control and Prevention; ECASS-4: European Cooperative Acute Stroke Study-4; NINDS: National Institute of Neurological Disorders and Stroke.



Documentation forms were collected from different contexts to provide a broad landscape of clinical care and clinical research documentation. The process of form collection and ensuring broad coverage by different documentation contexts is based on an established approach for CDE generation in the field of acute myeloid leukemia [10] and acute coronary syndrome [11]. Core contexts include routine clinical documentation, trial and register documentation, and international standards. To regard stroke-specific documentation, the context *rehabilitation* was added, as this was considered a highly relevant part of follow-up stroke care.

The core contexts of this work as follows:

1. Clinical routine documentation (stroke units): For this work the clinical routine documentation of the University Hospital Münster as a tertiary care hospital and the Elisabeth Hospital Gütersloh as an acute care hospital is considered, as well as the documentation from a tertiary care hospital in France. All forms were selected and made available by specialists working at these hospitals.
2. Documentation from rehabilitation: this category contains the clinical routine documentation from a facility for early rehabilitation in Gütersloh, a rehabilitation clinic in Meerbusch, a registry for early [19] and general rehabilitation [20] and also trial documentation from the University Hospital Halle [21].
3. Registries and quality standards: Two data models from the association of German stroke registries (ADSR) are considered, one regarding acute care for stroke patients in hospitals and one regarding thrombectomy and lysis therapy after stroke [9,22]. Moreover, a registry from Austria was included in the analysis [18].
4. Documentation from clinical trial research: The University Hospital Heidelberg provided the documentation for two clinical trials, a phase 3 trial researching lysis therapy for stroke patients (parameters: randomized, multicentre, double-blind, and placebo-controlled) and a trial researching decompressive surgery after severe stroke (parameters: prospective, multicentre, randomized, open, and controlled) [23,24].
5. International standards: The NINDS in the United States provides a list of the 50 most CDEs collected mainly from clinical trial documentation and guidelines [25]. Furthermore, the Centers for Disease Control and Prevention (CDC) provides performance measures for the quality of care–both of these data sets are considered in this work [26].

Processing of Documentation Forms and Semantic Coding

The collected forms were transformed into a unified document structure, called the operational data model (ODM) by the Clinical Data Interchange Standards Consortium. To do this, forms were created using the ODM editor available on the Medical Data-Models portal [12]. This portal is a metadata registry operated by the Institute of Medical Informatics in Münster and can be used to create, analyze, share, and reuse medical forms; it serves as an infrastructure for academic medical research (noncommercial) [16]. Each data element in the ODM format was assigned a Unified Medical Language System (UMLS) concept code. The UMLS Metathesaurus [27] contains UMLS concepts and incorporates important terminology, classification, and coding standards. For instance, the data element *birth date of patient* was allocated to the concept code *C0421451 Patient Date of Birth*.

Analysis and Generation of CDEs

UMLS-encoded ODM forms were analyzed using the CDE generator to identify common concepts and to generate CDEs [28]. The CDE generator is a publicly accessible tool and was used to count and display assigned UMLS codes ordered by frequency. It also enables the generation of cumulative concept coverage and pairwise comparisons of different documentation contexts. A concept was selected as a common concept if it appeared in at least two documentation contexts. By adding the most common information on datatypes, and permissible values from the analyzed sources, the corresponding CDEs were generated. A similar methodology was applied to other disease domains before [10,11].

Some concepts had to be aggregated and corrected to reveal overlaps between domains, for example, the concept *C0525032 International Normalized Ratio* and the concept *C0030605 Activated Partial Thromboplastin Time Measurement* were subsumed under the concept *C0005790 Blood Coagulation Tests* as part of the code harmonization. All forms and UMLS codes were checked by an experienced UMLS coder, and the resulting list of CDEs was also reviewed by a neurologist and a stroke specialist.

Finally, a list of the most frequented CDEs can be generated. In the following, the top-30 CDE extract shows the 30 most frequent concepts, which were used in at least two documentation contexts. The resulting list is shared on the Medical Data-Models portal for reuse.

Results

Data Collection

A total of 3287 data elements were identified based on 15 medical information management systems (Figure 2), from which 1051 unique medical concepts emerged.

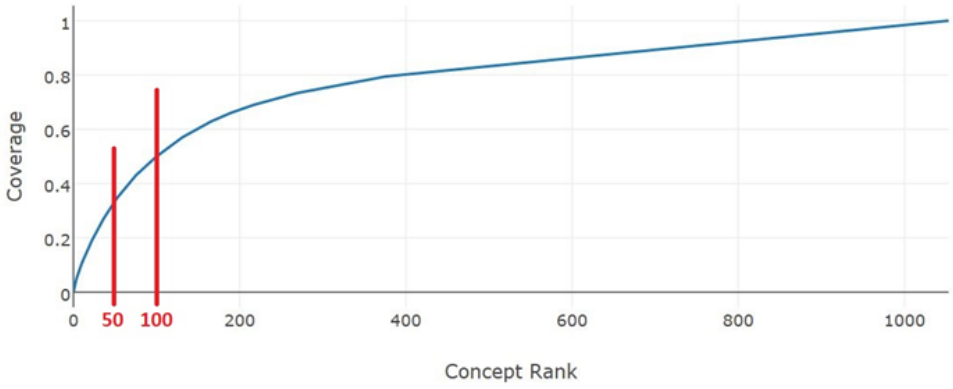
UMLS Coverage and Missing Concepts

Some data elements could not be assigned to a UMLS concept code in an unambiguous way; for instance, individual answers to intricate questions or complex instructions that were simply ticked off in a form (eg, *dose: 0.9 mg × kg body weight, maximum dose 90 mg, 10% initially as a bolus, the rest over 60 minutes via a syringe pump*). Moreover, administrative items such as dates and signatures appearing in medical reports were excluded from the concept code allocation. This study mainly focuses on medically relevant concepts.

Cumulative Concept Coverage

Cumulative frequencies help assess the heterogeneity of concepts. Among 1051 unique medical concepts, the 50 most frequent ones accounted for 34% of all concept occurrences in the collected stroke documentation (Figure 3). Expanding this to the 100 most frequent concepts leads to a coverage rate of 50%. The most frequent concept, *C0031809 Physical Examination*, refers to the general physical examination including the examination of the liver, the kidneys, the lungs, etc. Another frequently appearing concept (with a frequency of 43) was *C0003280 Anticoagulants* containing thrombosis prophylaxis and therapy, the dosing of heparin and vitamin K antagonists, etc; the concept *C3702515 Evaluation of Speech*, including examinations concerning articulation disorders, dysarthria, and apraxia, also appeared often. After the 363 most frequent concepts, the subsequent ones appeared only once. A complete list of concepts is provided in Multimedia Appendix 1.

Figure 3. Cumulative concept coverage.



Comparison of Documentation Contexts

Comparison of Clinical Routine Documentation and Documentation from Registries and Quality Standards (Context A and C)

Overall, routine clinical documentation of the stroke units contained 537 unique medical concepts and 206 concepts in registries and standards. Afterwards, the domains could be compared with the CDE generator. There was an overlap of 100 concepts between the information managements of the three stroke units and the information managements of the registries and standards (3 as well, Table 1); therefore, 100 concepts appeared in both domains, for example, concepts like *C0543414 Tobacco Use* or *C0007465 Cause of Death*. In contrast, concepts such as *C0018810 Heart Rate* only appeared in the information managements of the stroke units and *C0162578 Thrombectomy* only appeared in the information management of the registries and standards.

Table 1. Overview of overlaps between different contexts[a].

Contexts	Context A: clinical routine documentation (537 distinct concepts)	Context B: rehabilitation (475 distinct concepts)	Context C: registries and quality standards (206 distinct concepts)	Context D: clinical trial research (181 distinct concepts)	Context E: international standards (56 distinct concepts)
Context A: clinical routine documentation (537 distinct concepts)	*537*/100.0%/100.0%	*147*/27.4%/30.9%	*100*/18.6%/48,5%	*84*/15.6%/46.4%	*35*/6.5%/62.5%
Context B: rehabilitation (475 distinct concepts)	*147*/30.9%/27.4%	*475*/100.0%/100,0%	*85*/17.9%/41.3%	*52*/10.9%/28.7%	*28*/5.9%/50.0%
Context C: registries and quality standards (206 distinct concepts)	*100*/48.5%/18.6%	*85*/41.3%/17.9%	*206*/100.0%/100.0%	*56*/27.2%/30.9%	*25*/12.1%/44.6%
Context D: clinical trial research (181 distinct concepts)	*84*/46.4%/15.6%	*52*/28.7%/10.9%	*56*/30.9%/27.2%	*181*/100.0%/100,0%	*26*/14.4%/46.4%
Context E: international standards (56 distinct concepts)	*35*/62.5%/6.5%	*28*/50.0%/5.9%	*25*/44.6%/12.1%	*26*/46.4%/14.4%	*56*/100%/100%

[a]The first number (italicized) represents the number of common concepts between two contexts. For example, the second grid cell shows 147 shared concepts between context A and B, which corresponds to an overlap rate of 27.4% regarding context A (147/537) and 30.9% regarding context B (147/475).

Comparison of Clinical Routine Documentation and Documentation from Clinical Trial Research (Context A and D)

While the information management of the stroke units contained 537 unique medical concepts, the documentation from the clinical trial research (consisting of two trials, Table 1) yielded 181 unique medical concepts.

In total, 84 of these 181 concepts appeared in both the documentation from clinical trial research and clinical routine documentation, so there was an overlap of 46.4% between these domains. Examples of such overlapping concepts are *C0030193 Pain*, *C0005790 Blood Coagulation Tests*, and *C3476804*

NIHSS. Results from physical therapy (*C0949766 Physical Therapy*) were only recorded in the stroke units but did not appear in the clinical trials. By contrast, the concept *C3872877 Hemicraniotomy* was only found in the documentation from clinical trial research.

Comparison of National Registries and Standards (Germany) and International Standards (Context C and E)

Documentation from national registries and standards revealed 206 unique medical concepts, whereas only 56 concepts emerged from international standards. A total of 25 concepts appeared in both domains, for example *C0003811 Cardiac Arrhythmia*, *C0003280 Anticoagulants*, and *C0001948 Alcohol Consumption*. The concept *C2984908 Modified Rankin Scale* was only present in the German registries and standards and the concept *C0013658 Educational Status* only appeared in international standards.

International Standards and Documentation From Clinical Trial Research in Germany (Context E and D)

Documentation from two clinical trials conducted in Germany yielded 181 unique medical concepts, whereas the top list from the NINDS and the performance measures from the CDC contained 56 concepts. Of these 56 international standards concepts 26 (46%) overlapped with the concepts from the clinical trials, which represented 14.4% of the German trial documentation. Concepts like *C0005823 Blood Pressure*, *C2361123 Discharge Date*, and *C0040044 Thrombolytic Therapy* appeared in both domains, whereas the concept *C0009566 Complication* only appeared in the German clinical trials and the concept *C0001948 Alcohol Consumption* only appeared in the NINDS list of international standards.

Comparison of the Clinical Routine Documentation for Stroke Units and Rehabilitation Facilities (Domain A and Subdomain of B)

The routine documentation of the three stroke units contained 537 unique medical concepts, whereas the documentation of the rehabilitation clinics contained 475 unique medical concepts.

A comparison between these domains revealed an overlap of 147 concepts, for example, *C0011900 Diagnosis*, *C0154251 Lipid Metabolism Disorders*, and *C1305855 Body Mass Index*. The concept covering results from the ECG (*C0013798: Electrocardiogram*) was only present in the stroke unit documentation and *C0260682: Tracheostomy Status* and *C0233414: Disturbance of Attention* are examples of concepts that only appear in the documentation of the rehabilitation facilities.

Overview of CDEs

A list of the 50 most frequent concepts (based on the selection of a medical expert and medical computer scientists) could be created. Table 2 shows an extract containing the top 30 and sorted to different documentation categories. The absolute frequency of a concept could exceed the number sources (n=15), as some concepts appeared more than one in a source. A model implementation for an extended list of the 50 most common concepts was created as a Clinical Data Interchange Standards Consortium–compliant implementation for reuse (attached in the Multimedia Appendix 1).

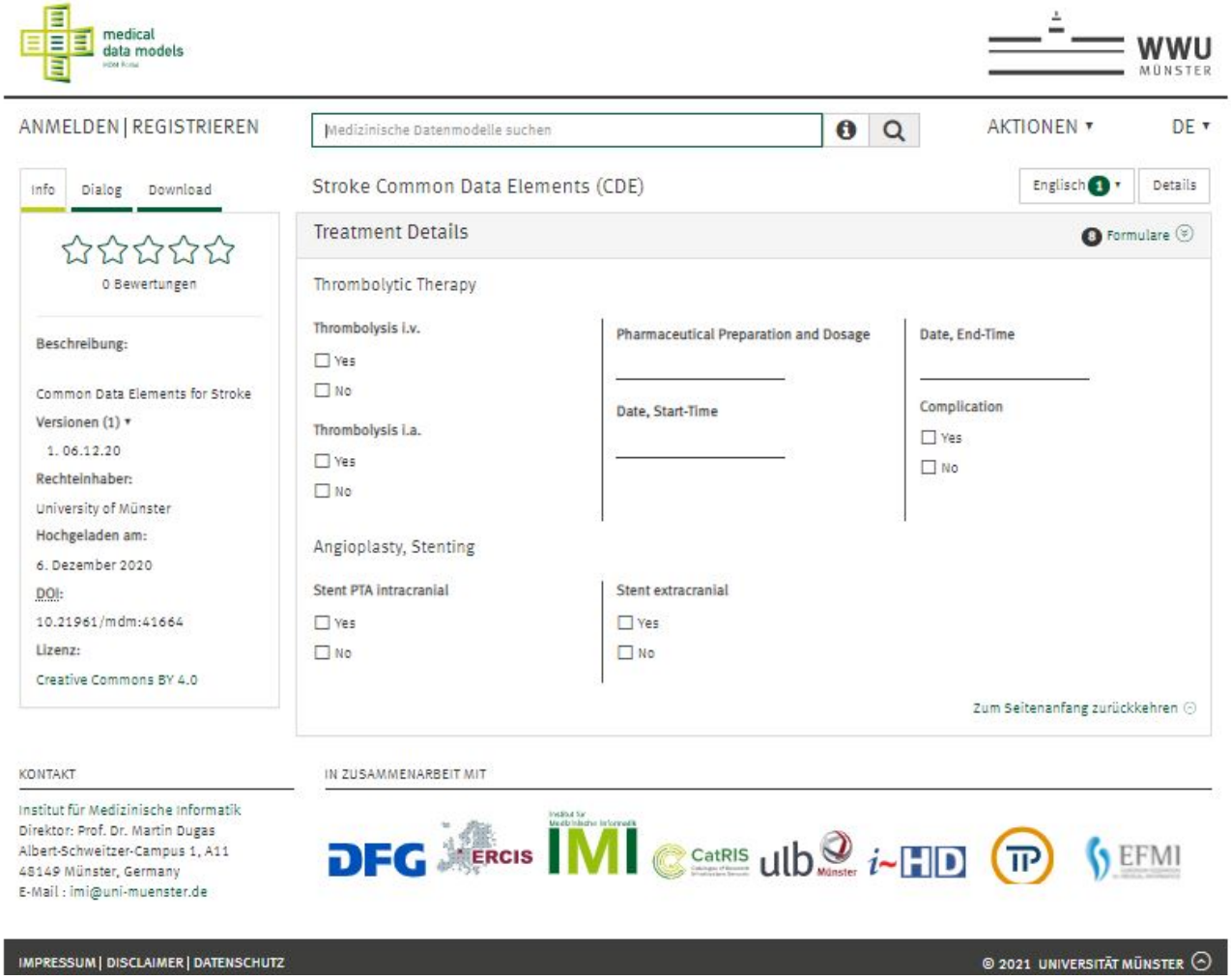
Moreover, the actual implementation according to the Clinical Data Interchange Standards Consortium ODM is available on the web [12] (Figure 4).

Table 2. Top 50 of most frequent concepts with absolute concept frequency and sorted by medical category[a].

No.	Concept	Documentation category	ACF	A	B	C	D	E
1	Patient internal identifier	Administrative/demographics	12	✓	✓	✓	✓	
2	Patient name	Administrative/demographics	9	✓	✓	✓	✓	
3	Gender	Administrative/demographics	9	✓	✓	✓	✓	✓
4	Patient date of birth, age	Administrative/demographics	19	✓	✓	✓	✓	✓
5	Date of admission	Administrative/demographics	13	✓	✓	✓	✓	✓
6	Discharge date	Administrative/demographics	12	✓	✓	✓	✓	✓
7	Living situation (alone, independent, or family home)	Administrative/demographics	17	✓	✓	✓	✓	
8	Death (finding, date/time)	Administrative/demographics	18	✓	✓	✓	✓	
9	Symptom findings in relation to time, time to treatment	Diagnostic/medical history	23	✓		✓	✓	✓
10	Diagnosis (transient ischemic attack, Ischemic stroke and localization, brain hemorrhage, and localization)	Diagnostic/medical history	132	✓	✓	✓	✓	✓
11	History of cerebrovascular accident and further medical history	Diagnostic/medical history	21	✓	✓	✓	✓	
12	Etiology	Diagnostic/medical history	19	✓		✓	✓	
13	Pre-existing conditions and risk factors	Diagnostic/medical history	104	✓	✓	✓	✓	✓
14	Vital signs (blood pressure, heart rate, SpO2, breathing rate, body temperature)	Diagnostic/medical history	67	✓	✓	✓	✓	✓
15	Neurologic symptoms and neurological deficit	Examination/follow-up	8	✓		✓		
16	General physical examination	Examination/follow-up	65	✓	✓	✓	✓	
17	National Institutes of Health Stroke Scale	Examination/follow-up	142	✓	✓	✓	✓	✓
18	Dysphagia/deglutition disorders	Examination/follow-up	23	✓	✓	✓		✓
19	Quality of vision-vision disorders?	Examination/follow-up	13	✓	✓			
20	Modified Rankin Scale	Nursing issues, rehabilitation	25	✓	✓	✓	✓	
21	Barthel index	Nursing issues, rehabilitation	17		✓	✓	✓	
22	Diagnostic imaging (magnetic resonance imaging, computed tomography, and ultrasonography)	Apparatus-based diagnostics	73	✓	✓	✓		
23	Angiography and digital subtraction	Apparatus-based diagnostics	9	✓	✓	✓		✓
24	Routine blood tests	Laboratory: blood panel	93	✓		✓	✓	
25	Medication list	Medication	37	✓	✓	✓	✓	✓
26	Anticoagulants	Medication	43	✓	✓	✓	✓	✓
27	Antiplatelet agents	Medication	30	✓	✓	✓	✓	✓
28	Thrombolytic therapy	Treatment details	23	✓	✓	✓	✓	✓
29	Angioplasty and stenting	Treatment details	26	✓		✓		
30	Physiotherapy and ergotherapy	Treatment details	38	✓	✓	✓		

[a]The common data elements (CDEs) for physiotherapy and ergotherapy are listed together in the 30th line. Columns, A, B, C, D and E present the occurrence of CDEs in the documentations of the according contexts (findings in Figure 2: Context A: Clinical routine documentation (stroke units), Context B: Documentation from rehabilitation, Context C: Registries and quality standards, Context D: Documentation from clinical trial research, and Context E: International standards).

Figure 4. The core data set in stroke care and research is available at the MDM-portal of the Institute of Medical Informatics in Münster [29].



Discussion

Principal Findings

This work systematically compared data from routine care, research, and quality management to derive a core data set, which could function as a starting point for further standardization efforts. The 50 most frequent of the 3287 concepts could already cover 34% of all concepts occurring in the entire data item collection. This shows the high potential of a potential core data set. The list has been published and exported in various documentation formats. This could facilitate data exchange between different institutions

During form collection, a few factors surfaced that made it a challenging endeavor: Issues related to intellectual property and confidentiality aspects impeded the availability of documentation forms, especially that of case report forms used in clinical trial research. Moreover, the software used in hospitals and rehabilitation facilities varied quite a bit, and often data were not available in digital form.

A general problem, especially with regard to routine documentation, is the numerous free-form text fields, which are only coded by an unspecific concept. In clinical routines, medical findings are often entered as free-text, whereas specific data points are usually queried in registries and standards and clinical trial research. The information captured by specific queries was also often found in free-form text fields in routine documentation; for example, a free-form text field asking for pre-existing conditions on the one hand and checkboxes for specific conditions (such as hypertension and atrial fibrillation) on the other.

We noticed a relatively small overlap of 25 concepts between national and international standards (comparison of national registries and standards [Germany] and International Standards [Context C and E]), which could be explained by the fact that German registries and standards focus mainly on routine documentation whereas the international standards of the NINDS on clinical trial research.

Clinical trials usually try to answer specific research questions (International standards and Documentation from Clinical Trial Research in Germany [Context E and D]). There was an overlap of 26 concepts between the international standards (represented by the top list of NINDS with the performance measures of CDC) and the documentations from clinical trials in Germany, but the overlaps were limited to basic patient data such as body weight, vital signs, and gender.

An interconnected and unified solution would facilitate the exchange and reuse of data, which would bring many benefits; for example, the automatic creation of text blocks for discharge

summaries would be possible. A more transparent documentation process would also simplify the efforts for standardization and quality assurance. Patient data from routine clinical documentation could be transferred to rehabilitation facilities, registries, and researchers conducting clinical trials. Data could also be reused, for instance, for further studies (secondary use), and it could easily be used for electronic health records.

This insight is not new, although earlier works and studies have already pointed out potential savings of both cost and time on the basis of CDEs for a more unified documentation process [6,7,11,30,31]. However, the implementation of CDEs and semantic annotations is not trivial. Some medical facilities and institutes would need to upgrade all their software and data entry forms that are currently still kept in paper form. There are many free-form text fields in current forms that are completed by individuals in a subjective manner. In this study, only the headers of free-form text fields were allocated to the concept codes. This led to important information not being covered, a smaller number of assigned concept codes, and thus, some overlaps are not recognizable.

UMLS is currently the most important approach for unifying the terminologies of biomedical resources, such as web-based databases and medical dictionaries. However, there are many UMLS codes that are quite similar and semantically nearly identical. Examples of such UMLS concept codes are *C2361123 Discharge Date*, *C2710998 Hospital Discharge Date*, and *C2361122 Discharge Date:Time Stamp--Date and Time:Point in Time:^Patient:Quantitative*. The coder has to decide which one to use, so encoding is a time-consuming process that is ideally performed by an experienced coder. A medical expert with experience in coding reviewed the concept of encoding for this work. The problem of uniform quality assurance for the assignment of CDEs was described previously [32,33].

Conclusions

Standardizing data from medical information management systems is necessary to reduce the amount of work needed for patient documentation and to allow for efficient querying, transfer, and reuse of data. Currently, there are no uniform standards for data collection regarding stroke across different domains. It would be strongly advisable to create a committee or work group to harmonize research and care relevant documentation for effective data reuse. This work provides a list of harmonized common data items based on existing stroke-related documentation, which can be reused to harmonize future documentation efforts in stroke-related care or research.

Acknowledgments

The authors would like to thank Heike Sieker and Dr Med Thomas Kloß, Elisabeth-Hospital Gütersloh, Germany; Dr Jasmin Heiss, TFS Trial Form Support GmbH c/o Business Center Bavaria, Munich, Germany, and Prof Dr Med Dipl Inf (FH) Peter A Ringleb, University Heidelberg, Germany; Prof Dr Eric Jüttler, University Heidelberg, Germany; Univ Prof Dr med Klaus Berger, Institute of Epidemiology and Social Medicine, University of Münster, Germany, Association of German Stroke Registries; Mag (FH) Alexander Gollmer, Health Austria GmbH, Vienna, Austria; Dr Med Volker Böhme, Landschaftsverband Westfalen-Lippe-Clinic Gütersloh, Germany; Prof Dr Med Stefan Knecht and Herrn Schicks, St. Mauritius Therapy Clinic, Meerbusch, Germany; Dr Med Björn Misselwitz (MPH), Business Office Quality Assurance Hessen, Eschborn, Germany; and Dr Rer Med Susanne Saal, Institute for Health and Health Care Sciences, Martin Luther University Halle-Wittenberg, Halle/Saale, Germany.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A complete list of concepts.
[XLSX File (Microsoft Excel File), 224 KB - medinform_v9i10e27396_app1.xlsx]

References

1. Fehr A, Lange C, Fuchs J, Neuhauser H, Schmitz R. Gesundheitsmonitoring und Gesundheitsindikatoren in Europa. J Health Monitor 2017;2(1):- [FREE Full text] [doi: 10.17886/RKI-GBE-2017-004.2]
2. Luengo-Fernandez R, Violato M, Candio P, Leal J. Economic burden of stroke across Europe: a population-based cost analysis. Eur Stroke J 2020 Mar;5(1):17-25 [FREE Full text] [doi: 10.1177/2396987319883160] [Medline: 32232166]
3. Stroke units. German Stroke Society. URL: https://www.dsg-info.de/stroke-units/stroke-units-uebersicht.html [accessed 2021-08-19]
4. Eckstein HH, Kühnl A, Berkefeld J, Dörfler A, Kopp I, Langhoff R, et al. Guideline detail view: diagnosis, therapy and follow-up care of extracranial carotid stenosis. AWMF Online. URL: https://www.awmf.org/leitlinien/detail/ll/004-028.html [accessed 2021-08-19]
5. Ammenwerth E, Spötl HP. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. Methods Inf Med 2009;48(1):84-91. [Medline: 19151888]

6. Bruland P, McGilchrist M, Zapletal E, Acosta D, Proeve J, Askin S, et al. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. BMC Med Res Methodol 2016 Nov 22;16(1):159 [FREE Full text] [doi: 10.1186/s12874-016-0259-3] [Medline: 27875988]
7. Damani R, Mayer S, Dhar R, Martin RH, Nyquist P, Olson DM, Unruptured Intracranial AneurysmsSAH CDE Project Investigators. Common data element for unruptured intracranial aneurysm and subarachnoid hemorrhage: recommendations from assessments and clinical examination workgroup/subcommittee. Neurocrit Care 2019 Jun;30(Suppl 1):28-35. [doi: 10.1007/s12028-019-00736-1] [Medline: 31090013]
8. Project support ZAV. German Stroke Society. URL: https://www.dsg-info.de/ [accessed 2021-08-19]
9. Current. Arbeitsgemeinschaft Deutschsprachiger Schlaganfall-Register. URL: https://www.schlaganfallregister.org/ [accessed 2021-08-19]
10. Holz C, Kessler T, Dugas M, Varghese J. Core data elements in acute myeloid leukemia: a unified medical language system-based semantic analysis and experts' review. JMIR Med Inform 2019 Aug 12;7(3):e13554 [FREE Full text] [doi: 10.2196/13554] [Medline: 31407666]
11. Kentgen M, Varghese J, Samol A, Waltenberger J, Dugas M. Common data elements for acute coronary syndrome: analysis based on the unified medical language system. JMIR Med Inform 2019 Aug 23;7(3):e14107 [FREE Full text] [doi: 10.2196/14107] [Medline: 31444871]
12. Dugas M, Hegselmann S, Riepenhausen S, Neuhaus P, Greulich L, Meidt A, et al. Compatible data models at design stage of medical information systems: leveraging related data elements from the MDM portal. Stud Health Technol Inform 2019 Aug 21;264:113-117. [doi: 10.3233/SHTI190194] [Medline: 31437896]
13. Schiariti V, Fowler E, Brandenburg JE, Levey E, Mcintyre S, Sukal-Moulton T, et al. A common data language for clinical research studies: the National Institute of Neurological Disorders and Stroke and American Academy for Cerebral Palsy and Developmental Medicine Cerebral Palsy Common Data Elements Version 1.0 recommendations. Dev Med Child Neurol 2018 Oct;60(10):976-986 [FREE Full text] [doi: 10.1111/dmcn.13723] [Medline: 29542813]
14. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. Clin Trials 2016 Dec;13(6):671-676 [FREE Full text] [doi: 10.1177/1740774516653238] [Medline: 27311638]
15. Grinnon ST, Miller K, Marler JR, Lu Y, Stout A, Odenkirchen J, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project - approach and methods. Clin Trials 2012 Jun;9(3):322-329 [FREE Full text] [doi: 10.1177/1740774512438980] [Medline: 22371630]
16. Mulcahey MJ, Vogel LC, Sheikh M, Arango-Lasprilla JC, Augutis M, Garner E, et al. Recommendations for the National Institute for Neurologic Disorders and Stroke spinal cord injury common data elements for children and youth with SCI. Spinal Cord 2017 Apr;55(4):331-340. [doi: 10.1038/sc.2016.139] [Medline: 27845358]
17. Suarez JI, Sheikh MK, Macdonald RL, Amin-Hanjani S, Brown RD, de Oliveira Manoel AL, Unruptured Intracranial Aneurysms and SAH CDE Project Investigators. Common data elements for unruptured intracranial aneurysms and subarachnoid hemorrhage clinical research: a national institute for neurological disorders and stroke and national library of medicine project. Neurocrit Care 2019 Jun;30(Suppl 1):4-19. [doi: 10.1007/s12028-019-00723-6] [Medline: 31087257]
18. Status quo of Austrian stroke care. In: Federal Ministry for Social Affairs, Health, Care and Consumer Protection (BMSGPK). Vienna: Ministry of Social Affairs; 2020.
19. [Record of stroke early rehabilitation in Hessen]. Geschäftsstelle Qualitätssicherung Hessen. URL: https://www.gqhnet.de/leistungsbereiche/schlaganfall/sa_dokumentationsboegen/SA_FRUEHREHA_2015_V01_Bogen.pdf [accessed 2021-10-04]
20. [Quality assurance in stroke treatment in Hessen]. Geschäftsstelle Qualitätssicherung Hessen. URL: https://www.gqhnet.de/leistungsbereiche/schlaganfall/sa_dokumentationsboegen/rehabogen2014_endfassung.pdf [accessed 2021-10-04]
21. Sal S, Becker C, Herrmann G, Kuss O, Lorenz S, Müller T, et al. Participative Rehabilitation in Stroke Patients (PaReSiS). ClinicalTrials.gov. 2012. URL: https://clinicaltrials.gov/ct2/show/NCT00687869 [accessed 2021-09-08]
22. Rohde S, Weber W, Berlis A, Urbach H, Reimer P, Schramm P, German Society of interventional Radiologyminimal invasive Therapy. Acute endovascular stroke treatment in Germany in 2019 : results from a nationwide database. Clin Neuroradiol 2021 Mar;31(1):11-19 [FREE Full text] [doi: 10.1007/s00062-020-00989-w] [Medline: 33481050]
23. Amiri H, Bluhmki E, Bendszus M, Eschenfelder CC, Donnan GA, Leys D, et al. European Cooperative Acute Stroke Study-4: extending the time for thrombolysis in emergency neurological deficits ECASS-4: ExTEND. Int J Stroke 2016 Feb;11(2):260-267. [doi: 10.1177/1747493015620805] [Medline: 26783318]
24. Jüttler E, Bösel J, Amiri H, Schiller P, Limprecht R, Hacke W, DESTINY II Study Group. DESTINY II: DEcompressive Surgery for the Treatment of malignant INfarction of the middle cerebral arterY II. Int J Stroke 2011 Feb;6(1):79-86. [doi: 10.1111/j.1747-4949.2010.00544.x] [Medline: 21205246]
25. Summary of core/supplemental-highly recommended recommendations: Stroke CDEs. National Institute of Neurological Disorders and Stroke. URL: https://www.commondataelements.ninds.nih.gov/sites/nindscde/files/Doc/Stroke/CDEStartupResource_Stroke.pdf [accessed 2021-08-19]
26. Centers for Disease Control and Prevention. Use of a registry to improve acute stroke care--seven states, 2005-2009. MMWR Morb Mortal Wkly Rep 2011 Feb 25;60(7):206-210 [FREE Full text] [Medline: 21346707]

27. Unified medical language system terminology services. NIH National Library of Medicine. URL: https://uts.nlm.nih.gov/uts/umls/home [accessed 2021-09-14]
28. Varghese J, Fujarski M, Hegselmann S, Neuhaus P, Dugas M. CDEGenerator: an online platform to learn from existing data models to build model registries. Clin Epidemiol 2018 Aug 10;10:961-970 [FREE Full text] [doi: 10.2147/CLEP.S170075] [Medline: 30127646]
29. Stroke Common Data Elements (CDE). URL: https://medical-data-models.org/41664 [accessed 2020-12-21]
30. Simko LC, Chen L, Amtmann D, Gibran N, Herndon D, Kowalske K, et al. Challenges to the standardization of trauma data collection in burn, traumatic brain injury, spinal cord injury, and other trauma populations: a call for common data elements for acute and longitudinal trauma databases. Arch Phys Med Rehabil 2019 May;100(5):891-898. [doi: 10.1016/j.apmr.2018.10.004] [Medline: 31030731]
31. Hackenberg KA, Etminan N, Wintermark M, Meyers PM, Lanzino G, Rüfenacht D, Unruptured Intracranial Aneurysms and SAH CDE Project Investigators. Common data elements for radiological imaging of patients with subarachnoid hemorrhage: proposal of a multidisciplinary research group. Neurocrit Care 2019 Jun;30(Suppl 1):60-78. [doi: 10.1007/s12028-019-00728-1] [Medline: 31115823]
32. Jiang G, Solbrig HR, Prud'hommeaux E, Tao C, Weng C, Chute CG. Quality assurance of cancer study common data elements using a post-coordination approach. AMIA Annu Symp Proc 2015 Nov 5;2015:659-668 [FREE Full text] [Medline: 26958201]
33. Huser V, Amos L. Analyzing real-world use of research common data elements. AMIA Annu Symp Proc 2018 Dec 5;2018:602-608 [FREE Full text] [Medline: 30815101]

Abbreviations

ADSR: Arbeitsgemeinschaft Deutschsprachiger Schlaganfall-Register
CDC: Centers for Disease Control and Prevention
CDE: common data element
NINDS: National Institute of Neurological Disorders and Stroke
ODM: operational data model
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 20.03.21; peer-reviewed by X Jing; comments to author 05.06.21; revised version received 12.07.21; accepted 19.07.21; published 12.10.21.

Please cite as:
Berenspöhler S, Minnerup J, Dugas M, Varghese J
Common Data Elements for Meaningful Stroke Documentation in Routine Care and Clinical Research: Retrospective Data Analysis
JMIR Med Inform 2021;9(10):e27396
URL: https://medinform.jmir.org/2021/10/e27396
doi:10.2196/27396
PMID:34636733

Original Paper

Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study

Donghwan Yun[1,2], MD; Semin Cho[2], MD; Yong Chul Kim[2], MD, PhD; Dong Ki Kim[2], MD, PhD; Kook-Hwan Oh[2], MD, PhD; Kwon Wook Joo[2], MD, PhD; Yon Su Kim[1,2], MD, PhD; Seung Seok Han[1,2], MD, PhD

[1]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea

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

Corresponding Author:
Seung Seok Han, MD, PhD
Department of Biomedical Sciences
Seoul National University College of Medicine
103 Daehakro, Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 4785 ext 8095
Fax: 82 2 745 2264
Email: hansway80@gmail.com

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2021/11/e34411

Abstract

Background: Precise prediction of contrast media–induced acute kidney injury (CIAKI) is an important issue because of its relationship with poor outcomes.

Objective: Herein, we examined whether a deep learning algorithm could predict the risk of intravenous CIAKI better than other machine learning and logistic regression models in patients undergoing computed tomography (CT).

Methods: A total of 14,185 patients who were administered intravenous contrast media for CT at the preventive and monitoring facility in Seoul National University Hospital were reviewed. CIAKI was defined as an increase in serum creatinine of ≥0.3 mg/dL within 2 days or ≥50% within 7 days. Using both time-varying and time-invariant features, machine learning models, such as the recurrent neural network (RNN), light gradient boosting machine (LGM), extreme gradient boosting machine (XGB), random forest (RF), decision tree (DT), support vector machine (SVM), κ-nearest neighbors, and logistic regression, were developed using a training set, and their performance was compared using the area under the receiver operating characteristic curve (AUROC) in a test set.

Results: CIAKI developed in 261 cases (1.8%). The RNN model had the highest AUROC of 0.755 (0.708-0.802) for predicting CIAKI, which was superior to that obtained from other machine learning models. Although CIAKI was defined as an increase in serum creatinine of ≥0.5 mg/dL or ≥25% within 3 days, the highest performance was achieved in the RNN model with an AUROC of 0.716 (95% confidence interval [CI] 0.664-0.768). In feature ranking analysis, the albumin level was the most highly contributing factor to RNN performance, followed by time-varying kidney function.

Conclusions: Application of a deep learning algorithm improves the predictability of intravenous CIAKI after CT, representing a basis for future clinical alarming and preventive systems.

(JMIR Med Inform 2021;9(10):e27177) doi:10.2196/27177

KEYWORDS

acute kidney injury; artificial intelligence; contrast media; deep learning; machine learning; kidney injury; computed tomography

Introduction

Computed tomography (CT) using contrast media is necessary to clinically detect abnormalities, but the administration of contrast media can lead to acute kidney injury (known as contrast media–induced acute kidney injury [CIAKI]). This is a critical issue due to subsequent risk of irreversible kidney dysfunction and increased mortality [1]. This adverse relationship is more critical in intra-arterial administration of contrast media than in intravenous administration [2]. Nevertheless, frequent use of CT scanning with intravenous contrast media increases the risk of nephrotoxicity, which requires prophylaxis and monitoring of kidney functions [3]. Prediction of intravenous CIAKI after CT scanning may be clinically essential to prepare for intervention in advance, but most relevant studies have primarily focused on intra-arterial CIAKI [4]. Models generated in some studies have predicted intravenous CIAKI, but these models had limitations because model performance was evaluated using a training set (rather than a test set) [5-10], an updated definition of CIAKI was not used [5-12], a prophylaxis protocol was not described [5,10,11], cases with intra-arterial administration of contrast media were combined in the analysis of intravenous cases [6,9,10], and confounding factors were not sufficiently considered [6-10].

Deep learning algorithms have achieved successful prediction of patient outcomes [13,14], which will change the paradigm of clinical decision making from diagnosis to treatment. Among deep learning algorithms, the recurrent neural network (RNN) can learn and characterize a temporal data set. In the nephrology field, using a time-varying data set of kidney function and vital signs, the predictability of outcomes has improved, such as acute kidney injury [15] and intradialytic complications, which are better than other machine learning (eg, gradient boosting machine) [16] and discrete-time logistic regression [17] models. Precise prediction of intravenous CIAKI may be difficult because multiple conditions have interactive and complex effects on its risk, and heterogeneous features of patients along with fluctuating dynamics of kidney functions before CT scanning may also complicate precise prediction. Herein, we addressed whether an RNN model with a time-varying data set including kidney functions could predict the risk of intravenous CIAKI better than other machine learning or conventional scoring models.

Methods

Data Source and Study Patients

A total of 19,628 patients underwent CT scanning with intravenous administration of contrast media at the 1-day-care facility of the Seoul National University Hospital between February 2007 and January 2019. This facility was built for the purpose of monitoring and preventing CIAKI in patients at risk, such as those with reduced kidney function or comorbidities. During admission, patients received hydration with 500 mL of 0.9% saline before and after intravenous administration of contrast media and 1200 mg of *N*-acetylcysteine for 3 days [18,19]. Kidney function was subsequently monitored for 2-7 days after CT scanning. Patients aged less than 18 years (n=5), with end-stage kidney disease (n=335), and no information about serum creatinine levels 28 days before and 7 days after CT scanning (n=5103) were excluded. Accordingly, 14,185 cases were included in the analysis (Multimedia Appendix 1). The institutional review board of the National University Hospital approved the study design (no. H-1812-134-997), which was conducted in accordance with the principles of the Declaration of Helsinki.

Study Features and Outcomes

Baseline characteristics, such as age, sex, weight, height, comorbidities (eg, coronary artery disease, any cancer, liver cirrhosis, glomerulonephritis, kidney transplantation), protocol of CT scanning and volume of contrast media, vital signs (eg, systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, and body temperature), and medications (eg, β-blocker, calcium channel blocker, angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, hydrochlorothiazide, spironolactone, furosemide, statin, metformin, sodium-glucose cotransporter 2 inhibitor, dipeptidyl peptidase-4 inhibitor, other oral hypoglycemic agents, and insulin), were collected using the patients' electronic medical records. Vital signs were measured at the time of admission to the facility. Laboratory findings were measured up to 1 month before CT scanning, and variables such as white blood cell count, hemoglobin, hematocrit, platelet count, cholesterol, albumin, total bilirubin, alkaline phosphatase, aspartate transaminase, alanine transaminase, uric acid, blood urea nitrogen, glucose, calcium, phosphate, sodium, potassium, chloride, and bicarbonate were evaluated. The estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease Epidemiology Collaboration equation [20]. Time-varying features included serum creatinine, eGFR, and elapsed times before CT scanning, and time-invariant features included all the other features. The baseline characteristics are summarized in Table 1.

Table 1. Baseline characteristics.

Features	Total (n=14,185)	CIAKI[a] (n=261)	Non-CIAKI (n=13,924)	*P* value[b]
Age (years), mean (range)	67.5 (56.7-78.4)	65.2 (54.1-76.3)	67.6 (56.7-78.5)	<.001
Male, n (%)	10,952 (77.2)	195 (74.7)	10,757 (77.3)	.33
Body mass index (kg/m²), mean (range)	24.0 (20.7-27.3)	24.0 (20.4-27.6)	24.0 (20.7-27.3)	.94
Type of CT[c], n (%)				
Abdomen and pelvis	4360 (30.7)	73 (28.0)	4287 (30.8)	N/A[d]
Liver	3323 (23.4)	90 (34.5)	3233 (23.2)	N/A
Urogenital	1330 (9.4)	17 (6.5)	1313 (9.4)	N/A
Chest	1004 (7.1)	15 (5.7)	989 (7.1)	N/A
Others	4168 (29.4)	66 (25.3)	4102 (29.5)	N/A
Contrast media volume (mL), mean (range)	98.3 (82.1-114.6)	99.8 (81.5-118.1)	98.3 (82.1-114.6)	.01
Vital signs				
Systolic blood pressure (mmHg), median (IQR)	126 (116-138)	130 (117.5-141)	126 (116-138)	.002
Diastolic blood pressure (mmHg), median (IQR)	75 (68-83)	78 (70-83.5)	75 (68-83)	.01
Heart rate (/min), median (IQR)	68 (61-79)	73 (62-82)	68 (61-79)	<.001
Respiratory rate (/min), mean (range)	18.3 (17.5-19.2)	18.3 (17.4-19.1)	18.3 (17.5-19.2)	.33
Body temperature (°C), mean (range)	36.4 (36.1-36.7)	36.4 (36.1-36.8)	36.4 (36.1-36.7)	.12
Comorbidities, n (%)				
Diabetes mellitus	4870 (34.3)	126 (48.3)	4744 (34.1)	<.001
Hypertension	6896 (48.6)	136 (52.1)	6760 (48.5)	.26
Coronary arterial disease	1940 (13.7)	28 (10.7)	1912 (13.7)	.16
Cancer, any type	11514 (81.2)	220 (84.3)	11294 (81.1)	.19
Liver cirrhosis	2253 (15.9)	58 (22.2)	2195 (15.8)	.005
Glomerulonephritis	439 (3.1)	13 (5.0)	426 (3.1)	.08
Kidney transplantation recipient	224 (1.6)	2 (0.8)	222 (1.6)	.29
Medication, n (%)				
Antihypertensive agents	5464 (38.5)	112 (42.9)	5352 (38.4)	.14
Diuretics	1905 (13.4)	71 (27.2)	1834 (13.2)	<.001
Statins	2731 (19.3)	59 (22.6)	2672 (19.2)	.17
Hypoglycemic agents	2553 (18.0)	58 (22.2)	2495 (17.9)	.07
Blood findings				
Hemoglobin (g/dL), median (IQR)	12.2 (10.6-13.7)	11.15 (10.1-12.4)	12.2 (10.7-13.7)	<.001
Hematocrit (%), median (IQR)	36.8 (32.4-40.9)	33.6 (30.4-37.95)	36.8 (32.5-41.1)	<.001
Albumin (g/dL), median (IQR)	4.1 (3.8-4.3)	3.8 (3.5-4.2)	4.1 (3.8-4.3)	<.001
Blood urea nitrogen (mg/dL), median (IQR)	22 (17-27)	25 (19-35)	22 (17-27)	<.001
Creatinine (mg/dL), median (IQR)	1.44 (1.25-1.67)	1.58 (1.27-2.01)	1.44 (1.24-1.67)	<.001
eGFR[e] (mL/min/1.73 m²), median (IQR)	47.1 (38.9-56.1)	42.7 (30.4-54.3)	47.2 (38.9-56.1)	<.001

[a]CIAKI: contrast media–induced acute kidney injury.

[b]*P* values were derived from the chi-square tests for categorical variables and the Student *t*-test or the Mann-Whitney *U* test for continuous variables.

[c]CT: computed tomography.

[d]N/A: not applicable.

[e]eGFR: estimated glomerular filtration rate.

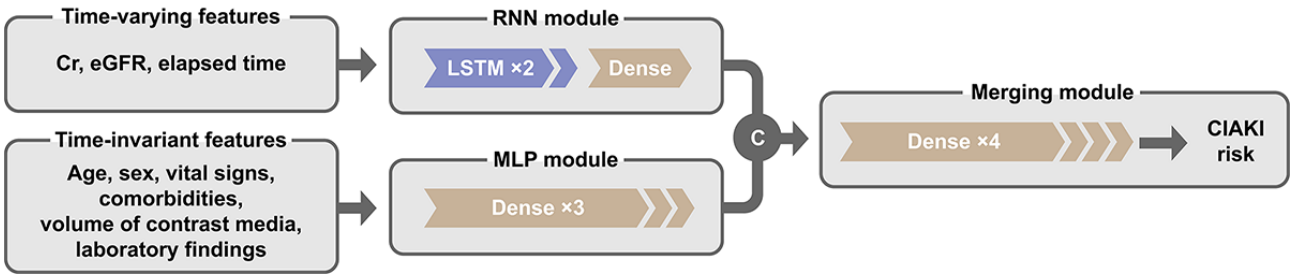
CIAKI was defined as an increase in serum creatinine of ≥0.3 mg/dL within 2 days or ≥50% within 7 days according to the Kidney Disease Improving Global Outcomes guideline [21]. In a sensitivity analysis, the other definition recommended by the European Society of Urogenital Radiology was used, such as an increase in serum creatinine of ≥0.5 mg/dL or ≥25% within 3 days [22]. As a long-term outcome, information about kidney progression (ie, doubling of serum creatinine, >50% decrease in eGFR, and the need for dialysis and transplantation) and all-cause mortality were obtained using the patients' electronic medical records, the Korean end-stage renal disease registry, and the National Database of Statistics, Korea.

Model Development

Patients were randomly assigned into a training set (70%) to develop the model and a test set (30%) to examine the performance of the model, wherein the occurrence of CIAKI was evenly distributed between the two sets. To develop the RNN model, we combined RNN and multiplayer perceptron (MLP) components. As an RNN component, we used the long short-term memory (LSTM) architecture, which is composed of input, output, and forget gates [23]. The median number of time-varying serum creatinine/eGFR values was 16 during the median timeframe of 4 years (1-9 years) before CT scanning. With respect to these results, 16 consecutive time-varying features were used in the RNN model. These features entered stacked cells and a subsequent dense layer (ie, RNN module), while time-invariant features were processed by 3 dense layers of the MLP module. The results were finally concatenated and then passed through 4 dense layers as a merging module. A dropout layer (rate=0.5) was followed behind each dense layer, while internal LSTM layers used input dropout (rate=0.5) and recurrent dropout (rate=0.5) [24]. Batch normalization layers were located at the end of RNN and multilayer perceptron modules and after the first and third layers of the merging module. Binary cross-entropy loss was used as a loss function to calculate the difference between actual and predicted labels. The Adam method was used for an optimizer [25], and the best parameter was selected using 10-fold cross-validation. Figure 1 presents the schematic diagram of the RNN model. To provide the model training process, we have added the Python code in Multimedia Appendix 2. The script includes data preprocessing, splitting, modeling, and training process information.

We also developed other machine learning models, such as a light gradient boosting machine (LGM), an extreme gradient boosting machine (XGB), a random forest (RF), a decision tree (DT), a support vector machine (SVM), a κ-nearest neighbor, and logistic regression, to compare their performance to the RNN model. These models could not handle time-varying features; therefore, only time-invariant features were included in the models. Tenfold cross-validation was used in the hyperparameter-tuning process, and candidate hyperparameters are listed in Multimedia Appendix 3.

Figure 1. Schematic diagram of the recurrent neural network. C: concatenate; CIAKI: contrast media–induced acute kidney injury; Cr: creatinine; Dense: dense layer; LSTM: long short-term memory; MLP: multilayer perceptron; eGFR: estimated glomerular filtration rate; RNN: recurrent neural network.



Feature Importance

Feature importance in the performance of the RNN model was evaluated using SHapley Additive exPlanations (SHAP) [26]. This method explains the model outcome as a sum of values attributed to each input feature, allowing the SHAP value to be interpreted as feature importance. The gradient SHAP model was applied to calculate the SHAP value [26]. The sum of SHAP values was used in the case of time-varying features. For non-RNN models, LinearExplainer (logistic regression and SVM) and TreeExplainer (DT, RF, XGB, and LGM) were used [26].

Statistical Analysis

Categorical and continuous variables are expressed as proportions and the means ± SD if they had a normal distribution and as medians with IQRs if they were non-normally distributed. Missing values of time-invariant features (4219 cases [28.5%] had at least 1 missing value) were imputed by the κ-nearest-neighboring imputer based on information in the training set [27]. If there were missing values in time-varying features (7031 cases [49.6%] had at least 1 missing value), masking was used during training of the RNN model. Model performance was evaluated in the test set using the area under the receiver operating characteristic curve (AUROC) and compared between models using the Delong test. All *P* values were set as two-sided, and values less than 0.05 were defined as significant. Statistical analyses were performed using R software (version 4.0.2; The Comprehensive R Archive Network: http://cran.r-project.org) and Python (version 3.8.3; Python Software Foundation: http://www.python.org). TensorFlow 2.3.0 (Google Brain, Google Inc.) was used as a deep learning framework [28], and other machine learning algorithms were performed by Scikit-learn [29].

Results

Baseline Characteristics

The mean age of cases was 67.5 (SD 11.1) years, and 22.8% (n=3233) were female. The median values of serum creatinine and eGFR were 1.4 mg/dL (IQR 1.3-1.7 mg/dL) and 47.1 mL/min/1.73 m² (IQR 38.9-56.1 mL/min/1.73 m²), respectively. The most common protocol was CT of the abdomen and pelvis (n=4360, 30.7%), followed by the liver (n=3323, 23.4%) and urogenital area (n=1330, 9.4%). Other baseline characteristics of the patients are presented in Table 1. The values of baseline characteristics did not differ between the training and test sets (Multimedia Appendix 4).

CIAKI and Long-Term Outcomes

Intravenous CIAKI occurred in 261 (1.8%) patients after CT scanning (1.8% in the training set and 2.0% in the test set). During the median follow-up period of 4 years (IQR 2-7 years), renal progression and all-cause mortality were identified in 3400 (24.0%) and 3762 (26.5%) patients, respectively. The CIAKI group had a higher risk of these outcomes compared with the non-CIAKI group (P<.001 for renal progression and P=.042 for all-cause mortality; see Multimedia Appendix 5).

Model Performance

When model performance was evaluated in the test set, the RNN model achieved the highest AUROC of 0.755 (95% confidence interval [CI] 0.708-0.802), followed by the RF (0.726 [95% CI 0.674-0.778]) and logistic regression (0.690 [95% CI 0.632-0.748]) (Table 2). The AUROC of the RNN model was greater than that obtained from other machine learning models (P<.05), except the RF, and the corresponding curves support these results (Figure 2).

Table 2. AUROC[a] of machine learning models in predicting intravenous CIAKI[b].

Models	AUROC (95% CI[c])	P value[d]
Logistic regression	0.690 (0.632-0.748)	.01
κ-Nearest neighbor	0.629 (0.566-0.693)	<.001
SVM[e]	0.644 (0.580-0.707)	<.001
DT[f]	0.633 (0.573-0.694)	<.001
RF[g]	0.726 (0.674-0.778)	.17
XGB[h]	0.665 (0.607-0.722)	.006
LGM[i]	0.651 (0.589-0.713)	<.001
RNN[j]	0.755 (0.708-0.802)	N/A[k]

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

[b]CIAKI: contrast media–induced acute kidney injury.

[c]CI: confidence interval.

[d]Compared to the receiver operating characteristic curve of the RNN model.

[e]SVM: support vector machine.

[f]DT: decision tree.

[g]RF: random forest.

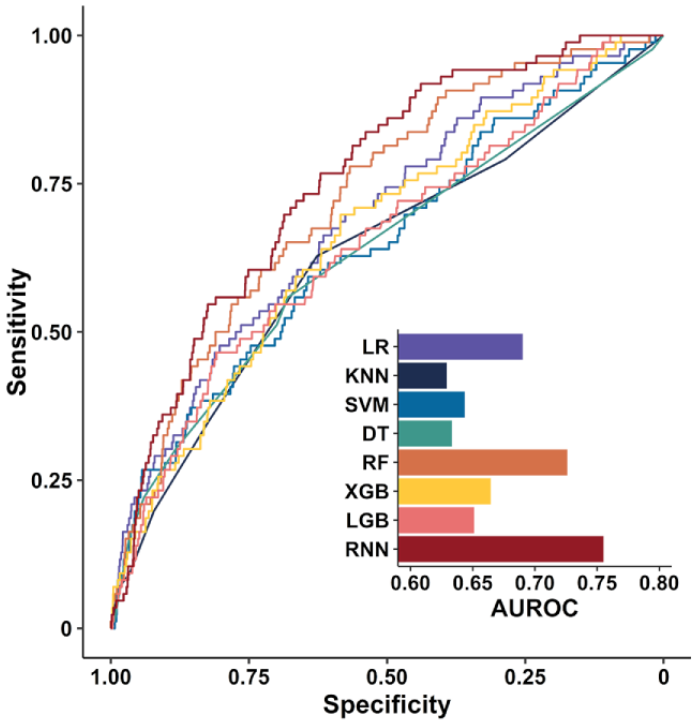
[h]XGB: extreme gradient boosting machine.

[i]LGM: light gradient boosting machine.

[j]RNN: recurrent neural network.

[k]N/A: not available.

Figure 2. AUROC for predicting intravenous CIAKI in the machine learning models. AUROC: area under the receiver operating characteristic curve; CIAKI: contrast media–induced acute kidney injury; DT: decision tree; KNN: κ-nearest neighbor; LGM: light gradient boosting machine; LR: logistic regression; SVM: support vector machine; RF: random forest; RNN: recurrent neural network; XGB: extreme gradient boosting machine.



We further compared the performance of the RNN model with other published scoring models. Eight studies have developed models to predict intravenous CIAKI [5-12]. The flowchart of study selection and their associated information is presented in Multimedia Appendix 6 and Table 3, respectively. Of these 8 models, 5 used specific features to develop models, such as cystatin C [6-8,10], homocysteine [7], neutrophil gelatinase-associated lipocalin [10], β2-microglobulin [10], and urine output [9]. Accordingly, 3 other models, such as the Mehran score [30], which was originally developed for patients undergoing intra-arterial administration of contrast media during coronary angiography but had also undergone CT scanning in 1 study [11], and two logistic regression–based models without testing of an independent data set [5,12], were compared to the RNN model. The performance of these 3 models was lower than that of the RNN model with the following AUROCs: 0.521 (P<.001) in the Mehran score and 0.539 (P<.001) and 0.645 (P=.022) in the other 2 logistic regression-based models.

Table 3. Previous studies predicting intravenous CIAKI[a].

Studies	Study subjects	CIAKI definition	CIAKI (%)	Prophylaxis protocol	Patients, n (training/test)	Features, n	Modeling methods	AUROC[b] in test set
Kim et al [5]	Abdominal CT[c] in emergency department	≥0.5 mg/dL or ≥25% within 3 days	4.5	Not declared	750/0	2	Nomogram	N/A[d] (0.794 in training set)
Wacker-Gussmann et al [6][d]	CAG[e] or CT in hospitalized patients with sCr[f] levels between 0.8 and 1.3 mg/dL	≥0.5 mg/dL or ≥25% within 48 h	14.2	Oral fluid intake, 2 L	373/0	2	Baseline ratio of CysC[g]/Cr	N/A (0.826 in training set)
Li et al [7]	Coronary CT in patients with eGFR[h] of ≥60 mL/min/1.73 m^2	≥0.5 mg/dL or ≥25% within 48 h	9.8	Oral fluid intake, 500 mL	580/0	5	AUROC with single feature	N/A (0.829 of homocysteine in training set)
Li et al [8]	Coronary CT in patients with eGFR of ≥60 mL/min/1.73 m^2	≥0.5 mg/dL or ≥25% within 48 h	12.3	Oral fluid intake, 500 mL	424/0	2	AUROC with single feature	N/A (0.781 of CysC in training set)
Hocine et al [9][i]	CAG or CT in intensive care unit	≥0.5 mg/dL or ≥25% within 3 days	60.1	No routine protocol	149/0	1	RIFLE[j] criteria	N/A
Ho et al [11]	CT pulmonary angiogram in intensive care unit	>0.5 mg/dL within 48 h	40.9	Not declared	0[a]/137	8	Mehran score	0.864
Jeon et al [12]	CT in cancer patients with eGFR of <45 mL/min/1.73 m^2	>25% within 2-6 days	2.46	0.9% Saline with *N*-acetylcysteine	2185/539	3	Scoring system based on logistic regression	0.749
Banda et al [10][k]	CAG and CT in hospitalized patients	>0.5 mg/dL or >25% within 48-72 h	N/A	Not declared	90/0	5	AUROC with single feature	N/A (0.684 of β2-microglobulin in training set)

[a]CIAKI: contrast media–induced acute kidney injury.

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

[c]CT: computed tomography.

[d]N/A: not available.

[e]CAG: coronary angiography.

[f]sCr: serum creatinine.

[g]CysC: cystatin C.

[h]eGFR: estimated glomerular filtration rate.

[i]Used the Mehran risk score.

[j]RIFLE: Risk Injury Failure Loss of kidney function and End-stage kidney disease classification.

[k]Included patients with both intravenous and intra-arterial administration of contrast media.

Sensitivity Analysis

For sensitivity analysis, another definition of CIAKI was used, an increase in serum creatinine of ≥0.5 mg/dL or ≥25% within 3 days [22]. The RNN model was the best model in predicting the risk of CIAKI, with an AUROC of 0.716 (95% CI 0.664-0.768), which was greater than that of most of the other machine learning models (Multimedia Appendix 7). The corresponding curves support these results (Multimedia Appendix 8).

Other machine learning models were trained after including 48 features (ie, 16 sets of serum creatinine, eGFR, and elapsed times) as an independent feature without timed order. The results are summarized in Multimedia Appendix 9. Although these features were considered in the models, the model performance was less than that of the RNN model.

Furthermore, the original pipeline was separated into 4 models (MLP alone, MLP plus merging, RNN alone, and RNN plus merging), and their performance was compared with that of the original pipeline (named a default model). The AUROC plots are presented in Multimedia Appendix 10. The deep learning model with the MLP module alone and the RNN module alone had AUROCs of 0.705 (95% CI 0.647-0.763) and 0.702 (95% CI 0.642-0.763), respectively. After adding the merging module to these models, the AUROCs were 0.710 (95% CI 0.653-0.768) in the MLP-plus-merging module and 0.675 (95% CI 0.610-0.740) in the RNN-plus-merging module. All these values were lower than the value from the original deep learning model.

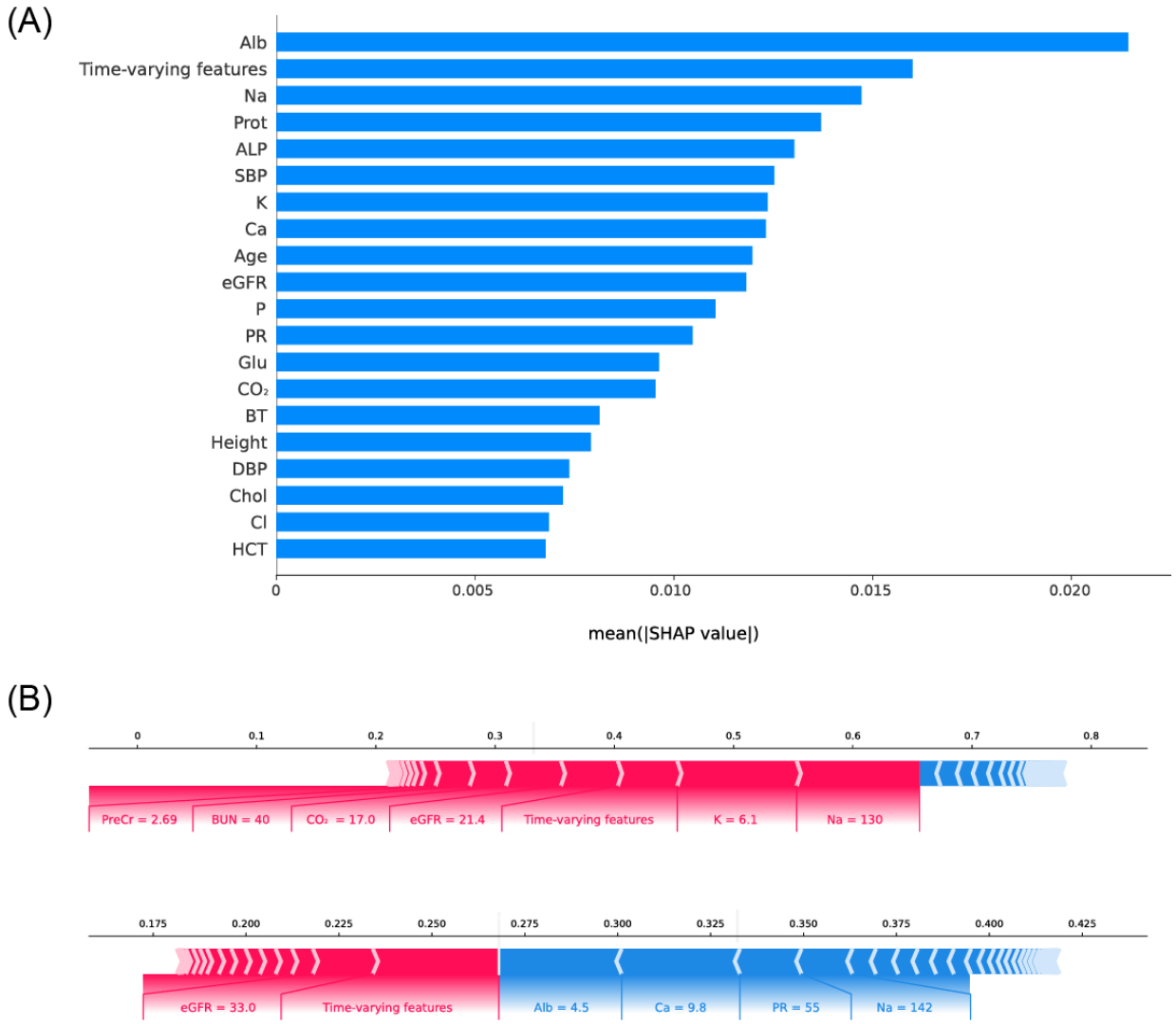
To evaluate the effect of the model complexity on performance, we built other deep learning architectures, such as a simple model (ie, 1 less dense layer in the RNN module, MLP module, and merging module) and a complex model (ie, 1 more dense layer in the RNN module, MLP module, and merging module). The AUROCs were 0.751 (95% CI 0.702-0.801) and 0.734 (95% CI 0.678-0.791) in the simple and complex models, respectively. We also developed models with a single LSTM layer having a simpler RNN architecture (named "single model") and with two stacked bidirectional LSTM layers having a more complex RNN architecture (named "bidirectional model"). The single and bidirectional models had AUROCs of 0.746 (95% CI 0.696-0.795) and 0.717 (95% CI 0.656-0.777), respectively. The AUROC plots of these models compared to that of the original model (named "default model") are described in Multimedia Appendix 11.

Feature-Ranking Analysis

Feature importance in RNN performance was estimated using SHAP (Figure 3A). Serum albumin had the highest impact on model output, and time-varying serum creatinine was ranked second. Age, several laboratory features (eg, sodium, protein, and alkaline phosphatase), and vital signs (eg, systolic blood pressure) were also highly ranked. We also explored SHAP values in non-RNN machine learning models (Multimedia Appendix 12). In the RF model and the LGM model, which achieved the second- and third-highest performance, SHAP values were highly correlated (Pearson's correlation of the mean of absolute SHAP values=0.781; $P<0.001$; Multimedia Appendix 13), and the time-invariant features with high impact in the RNN model (eg, albumin, sodium, and protein) were also highly ranked.

Figure 3B shows 2 representative cases with CIAKI. The model predicted the risk of CIAKI as 0.680 (true-positive) and 0.264 (false-positive) in the upper and lower cases, respectively. According to SHAP analysis, hyponatremia, hyperkalemia, time-varying serum features, and low eGFR contributed to precise prediction in the upper case. In the lower case, although serum albumin, calcium, and other parameters underestimated the risk of CIAKI, the time-varying features and low eGFR corrected this false prediction.

Figure 3. SHAP analysis of the RNN model. (A) Feature ranking according to SHAP value. (B) Two cases to explain the risk of intravenous CIAKI with SHAP values. RNN: recurrent neural network; SHAP: SHapley Additive eXplanations; CIAKI: contrast media–induced acute kidney injury; Alb: albumin; ALP: alkaline phosphatase; BT: body temperature; Ca: calcium; Chol: cholesterol; Cl: chloride; CO_2: bicarbonate; DBP: diastolic blood pressure; eGFR; estimated glomerular filtration rate; Glu: glucose; HCT: hematocrit: K: potassium; Na: sodium; P: phosphate; PR: pulse rate; PreCr; baseline creatinine; Prot: protein; SBP: systolic blood pressure.



Discussion

Principal Results

Intravenous CIAKI is a critical issue because it contributes to poor outcomes [31], as noted in its association with renal progression and increased mortality above. This study first applied the RNN algorithm to predict intravenous CIAKI with a greater AUROC than that obtained from other machine learning or conventional scoring models. These results indicate that the time-varying data of kidney function (ie, serum creatinine and eGFR) significantly contribute to the precise prediction of intravenous CIAKI. SHAP analysis demonstrated that feature importance could help understand how risk is estimated.

Because kidney function fluctuates over time, a single value of serum creatinine or eGFR may not perfectly represent the kidney function of patients. Certain attempts using time-varying kidney functions by time-dependent Cox regression [32] and trajectory analysis [33] have improved the precise estimation of kidney function. Recently, deep learning with the RNN model showed favorable performance in predicting acute kidney injury [15], implying the additive benefit of time-varying kidney functions to the model performance. Patients with comorbidities, including cancer, diabetes mellitus, and chronic kidney disease, are recommended for frequent follow-up of their kidney function because these data can be used to better predict the trend of kidney function than a single estimation. In this regard, the present RNN model achieved the highest performance in predicting intravenous CIAKI with time-varying features.

Deep learning architecture is complex and difficult to interpret in nature and is referred to as a black box. To overcome this limitation, this study applied SHAP to concretely explain the model output. Using SHAP values, clinicians can comprehend how the risk probability is explained by the results of various features and decide whether the model output is feasible. If the model prediction seems to be imprecise, as in the lower case in

Figure 3B, the SHAP values in features highly relevant to the model performance provide room for reconsideration.

Limitations

Despite these informative results, there are limitations to be discussed. The study design was retrospective and needs to be validated in future independent cohorts. Unidentified factors, such as urine output and heart function, may provide additional information about the risk of CIAKI, but the present data set included most clinically used features. The prophylaxis protocol may differ between centers, and thus, the present RNN model may need to be adjusted when applied externally.

Conclusions

Application of a deep learning algorithm improves the predictability of intravenous CIAKI, and our model performs better than other machine learning and conventional scoring models. These results may be attributable to the consideration of time-varying kidney functions, in addition to time-invariant features, and corresponding SHAP values may maximize the utility of the model in clinics. If proactive management of intravenous CIAKI is possible via precise prediction, overall patient outcomes will improve. The study results represent the basis of this goal.

Acknowledgments

The data sets used and analyzed in this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Flowchart of data retrieval and splitting.
[PNG File , 21 KB - medinform_v9i10e27177_app1.png]

Multimedia Appendix 2
Python pseudocode including data preprocessing, splitting, modeling, and training processes.
[TXT File , 9 KB - medinform_v9i10e27177_app2.txt]

Multimedia Appendix 3
Hyperparameters used in machine learning models.
[DOCX File , 22 KB - medinform_v9i10e27177_app3.docx]

Multimedia Appendix 4
Baseline characteristics in the training and test sets.
[DOCX File , 29 KB - medinform_v9i10e27177_app4.docx]

Multimedia Appendix 5
Kaplan-Meier curves of renal survival (A) and patient survival (B) according to intravenous CIAKI. CIAKI: contrast media–induced acute kidney injury.
[PNG File , 32 KB - medinform_v9i10e27177_app5.png]

Multimedia Appendix 6
Flowchart of study selection regarding the modeling of predicting intravenous CIAKI. CIAKI: contrast media–induced acute kidney injury.
[PNG File , 32 KB - medinform_v9i10e27177_app6.png]

Multimedia Appendix 7
Table of AUROCs for predicting intravenous CIAKI, which was defined as an increase in serum creatinine ≥0.5 mg/dL or ≥25% within 3 days. AUROC: area under the receiver operating characteristic curve; CIAKI: contrast media–induced acute kidney injury.
[DOCX File , 23 KB - medinform_v9i10e27177_app7.docx]

Multimedia Appendix 8
Plots showing AUROCs for predicting intravenous CIAKI, which was defined as an increase in serum creatinine ≥0.5 mg/dL or ≥25% within 3 days. AUROC: area under the receiver operating characteristic curve; CIAKI: contrast media–induced acute kidney injury.

[PNG File , 48 KB - medinform_v9i10e27177_app8.png]

Multimedia Appendix 9
AUROCs of machine learning models in predicting intravenous CIAKI. AUROC: area under the receiver operating characteristic curve; CIAKI: contrast media–induced acute kidney injury.
[DOCX File , 139 KB - medinform_v9i10e27177_app9.docx]

Multimedia Appendix 10
AUROC for predicting intravenous CIAKI according to the combination of modules. AUROC: area under the receiver operating characteristic curve CIAKI: contrast media–induced acute kidney injury.
[PNG File , 69 KB - medinform_v9i10e27177_app10.png]

Multimedia Appendix 11
AUROC for predicting intravenous CIAKI in RNN models. AUROC: area under the receiver operating characteristic curve; CIAKI: contrast media–induced acute kidney injury; RNN: recurrent neural network.
[PNG File , 70 KB - medinform_v9i10e27177_app11.png]

Multimedia Appendix 12
SHAP analysis of machine learning models. SHAP: SHapley Additive exPlanations.
[DOCX File , 214 KB - medinform_v9i10e27177_app12.docx]

Multimedia Appendix 13
Scattered plot showing the paired mean of absolute SHAP values of 53 time-invariant features in random forest and LGM models. SHAP: SHapley Additive exPlanations; LGM: light gradient boosting machine.
[PNG File , 87 KB - medinform_v9i10e27177_app13.png]

References

1. Rudnick MR, Leonberg-Yoo AK, Litt HI, Cohen RM, Hilton S, Reese PP. The controversy of contrast-induced nephropathy with intravenous contrast: what is the risk? Am J Kidney Dis 2020 Jan;75(1):105-113. [doi: 10.1053/j.ajkd.2019.05.022] [Medline: 31473019]
2. Mehran R, Dangas GD, Weisbord SD. Contrast-associated acute kidney injury. N Engl J Med 2019 May 30;380(22):2146-2155. [doi: 10.1056/nejmra1805256]
3. Thurley P, Crookdake J, Norwood M, Sturrock N, Fogarty AW. Demand for CT scans increases during transition from paediatric to adult care: an observational study from 2009 to 2015. Br J Radiol 2018 Feb;91(1083):20170467 [FREE Full text] [doi: 10.1259/bjr.20170467] [Medline: 29144163]
4. Silver SA, Shah PM, Chertow GM, Harel S, Wald R, Harel Z. Risk prediction models for contrast induced nephropathy: systematic review. BMJ 2015 Aug 27;351:h4395 [FREE Full text] [doi: 10.1136/bmj.h4395] [Medline: 26316642]
5. Kim KS, Kim K, Hwang SS, Jo YH, Lee CC, Kim TY, et al. Risk stratification nomogram for nephropathy after abdominal contrast-enhanced computed tomography. Am J Emerg Med 2011 May;29(4):412-417. [doi: 10.1016/j.ajem.2009.11.015] [Medline: 20825813]
6. Wacker-Gußmann A, Bühren K, Schultheiss C, Braun SL, Page S, Saugel B, et al. Prediction of contrast-induced nephropathy in patients with serum creatinine levels in the upper normal range by cystatin C: a prospective study in 374 patients. AJR Am J Roentgenol 2014 Feb;202(2):452-458. [doi: 10.2214/AJR.13.10688] [Medline: 24450691]
7. Li S, Tang X, Peng L, Luo Y, Zhao Y, Chen L, et al. A head-to-head comparison of homocysteine and cystatin C as pre-procedure predictors for contrast-induced nephropathy in patients undergoing coronary computed tomography angiography. Clin Chim Acta 2015 Apr 15;444:86-91. [doi: 10.1016/j.cca.2015.02.019] [Medline: 25687162]
8. Li S, Zheng Z, Tang X, Peng L, Luo Y, Dong R, et al. Preprocedure and postprocedure predictive values of serum β2-microglobulin for contrast-induced nephropathy in patients undergoing coronary computed tomography angiography: a comparison with creatinine-based parameters and cystatin C. J Comput Assist Tomogr 2015;39(6):969-974. [doi: 10.1097/RCT.0000000000000294] [Medline: 26248154]
9. Hocine A, Defrance P, Lalmand J, Delcour C, Biston P, Piagnerelli M. Predictive value of the RIFLE urine output criteria on contrast-induced nephropathy in critically ill patients. BMC Nephrol 2016 Mar 28;17:36 [FREE Full text] [doi: 10.1186/s12882-016-0243-5] [Medline: 27021438]
10. Banda J, Duarte R, Dix-Peek T, Dickens C, Manga P, Naicker S. Biomarkers for diagnosis and prediction of outcomes in contrast-induced nephropathy. Int J Nephrol 2020;2020:8568139 [FREE Full text] [doi: 10.1155/2020/8568139] [Medline: 32411464]

11. Ho KM, Harahsheh Y. Predicting contrast-induced nephropathy after CT pulmonary angiography in the critically ill: a retrospective cohort study. J Intensive Care 2018;6:3 [FREE Full text] [doi: 10.1186/s40560-018-0274-z] [Medline: 29387419]
12. Jeon J, Kim S, Yoo H, Kim K, Kim Y, Park S, et al. Risk prediction for contrast-induced nephropathy in cancer patients undergoing computed tomography under preventive measures. J Oncol 2019;2019:8736163 [FREE Full text] [doi: 10.1155/2019/8736163] [Medline: 31057617]
13. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. The Lancet Respiratory Medicine 2018 Dec;6(12):905-914. [doi: 10.1016/S2213-2600(18)30300-X] [Medline: 30274956]
14. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
15. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019 Aug;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
16. Koyner JL, Carey KA, Edelson DP, Churpek MM. The development of a machine learning inpatient acute kidney injury prediction model*. Crit Care Med 2018;46(7):1070-1077. [doi: 10.1097/ccm.0000000000003123]
17. Simonov M, Ugwuowo U, Moreira E, Yamamoto Y, Biswas A, Martin M, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: a descriptive modeling study. PLoS Med 2019 Jul 15;16(7):e1002861 [FREE Full text] [doi: 10.1371/journal.pmed.1002861] [Medline: 31306408]
18. Marenzi G, Assanelli E, Marana I, Lauri G, Campodonico J, Grazi M, et al. N-Acetylcysteine and Contrast-Induced Nephropathy in Primary Angioplasty. N Engl J Med 2006 Jun 29;354(26):2773-2782. [doi: 10.1056/nejmoa054209]
19. Fishbane S, Durham JH, Marzo K, Rudnick M. N-acetylcysteine in the prevention of radiocontrast-induced nephropathy. J Am Soc Nephrol 2004 Feb 01;15(2):251-260 [FREE Full text] [doi: 10.1097/01.asn.0000107562.68920.92] [Medline: 14747371]
20. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF, Feldman HI, Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI). A new equation to estimate glomerular filtration rate. Ann Intern Med 2009 May 05;150(9):604-612 [FREE Full text] [doi: 10.7326/0003-4819-150-9-200905050-00006] [Medline: 19414839]
21. 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]
22. Morcos SK, Thomsen HS, Webb JA. Contrast-media-induced nephrotoxicity: a consensus report. Contrast Media Safety Committee, European Society of Urogenital Radiology (ESUR). Eur Radiol 1999;9(8):1602-1613. [doi: 10.1007/s003300050894] [Medline: 10525875]
23. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. ArXiv 2015 May 20;52(10):52-5098-52-5098 [FREE Full text] [doi: 10.5860/choice.189890]
24. Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks. Adv Neural Inf Process Syst 2016:1019-1027.
25. Kingma D, Ba J. Adam: A method for stochastic optimization. ArXiv 2014:A.
26. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems 2017:4768-4777.
27. Batista G, Monard M. A study of k-nearest neighbour as an imputation method. HIS 2002;87:48.
28. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J. Tensorflow: a system for large-scale machine learning. 2016 Nov 2 Presented at: 12th USENIX Symposium on Operating Systems Design and Implementation; 2016 Nov 2–4; Savannah, GA, USA p. 265-283.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit learn: machine learning in Python. J Mach Learn Res 2011:2825-2830 [FREE Full text]
30. Mehran R, Aymong E, Nikolsky E, Lasic Z, Iakovou I, Fahy M, et al. A simple risk score for prediction of contrast-induced nephropathy after percutaneous coronary intervention: development and initial validation. J Am Coll Cardiol 2004 Oct 06;44(7):1393-1399. [doi: 10.1016/s0735-1097(04)01445-7]
31. Maioli M, Toso A, Leoncini M, Gallopin M, Musilli N, Bellandi F. Persistent renal damage after contrast-induced acute kidney injury: incidence, evolution, risk factors, and prognosis. Circulation 2012 Jun 26;125(25):3099-3107. [doi: 10.1161/CIRCULATIONAHA.111.085290] [Medline: 22592896]
32. Dekker FW, de Mutsert R, van Dijk PC, Zoccali C, Jager KJ. Survival analysis: time-dependent effects and time-varying risk factors. Kidney Int 2008 Oct;74(8):994-997 [FREE Full text] [doi: 10.1038/ki.2008.328] [Medline: 18633346]
33. Kang E, Han SS, Kim J, Park SK, Chung W, Oh YK, et al. Discrepant glomerular filtration rate trends from creatinine and cystatin C in patients with chronic kidney disease: results from the KNOW-CKD cohort. BMC Nephrol 2020 Jul 16;21(1):280-289 [FREE Full text] [doi: 10.1186/s12882-020-01932-4] [Medline: 32677901]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CI: confidence interval
CIAKI: contrast media–induced acute kidney injury
CT: computed tomography
DT: decision tree
eGFR: estimated glomerular filtration rate
LGM: light gradient boosting machine
LSTM: long short-term memory
MLP: multiplayer perceptron
RF: random forest
RNN: recurrent neural network
SHAP: SHappley Additive exPlanations
SVM: support vector machine
XGB: extreme gradient boosting machine

Edited by C Lovis, J Hefner; submitted 18.01.21; peer-reviewed by A Staffini, G Lim, JA Benítez-Andrades; comments to author 08.03.21; revised version received 05.04.21; accepted 03.09.21; published 01.10.21.

Please cite as:
Yun D, Cho S, Kim YC, Kim DK, Oh KH, Joo KW, Kim YS, Han SS
Use of Deep Learning to Predict Acute Kidney Injury After Intravenous Contrast Media Administration: Prediction Model Development Study
JMIR Med Inform 2021;9(10):e27177
URL: https://medinform.jmir.org/2021/10/e27177
doi:10.2196/27177
PMID:34596574

Original Paper

Predicting the Linguistic Accessibility of Chinese Health Translations: Machine Learning Algorithm Development

Meng Ji[1], PhD; Pierrette Bouillon[2], PhD

[1]School of Languages and Cultures, University of Sydney, Sydney, Australia

[2]University of Geneva, Geneva, Switzerland

Corresponding Author:
Meng Ji, PhD
School of Languages and Cultures
University of Sydney
City Road Camperdown/Darlington
Sydney, 2006
Australia
Phone: 61 04 3406 9975
Email: christine.ji@sydney.edu.au

Abstract

Background: Linguistic accessibility has an important impact on the reception and utilization of translated health resources among multicultural and multilingual populations. Linguistic understandability of health translation has been understudied.

Objective: Our study aimed to develop novel machine learning models for the study of the linguistic accessibility of health translations comparing Chinese translations of the World Health Organization health materials with original Chinese health resources developed by the Chinese health authorities.

Methods: Using natural language processing tools for the assessment of the readability of Chinese materials, we explored and compared the readability of Chinese health translations from the World Health Organization with original Chinese materials from the China Center for Disease Control and Prevention.

Results: A pairwise adjusted *t* test showed that the following 3 new machine learning models achieved statistically significant improvement over the baseline logistic regression in terms of area under the curve: C5.0 decision tree (95% CI –0.249 to –0.152; $P<0.001$), random forest (95% CI 0.139-0.239; $P<0.001$) and extreme gradient boosting tree (95% CI 0.099-0.193; $P<0.001$). There was, however, no significant difference between C5.0 decision tree and random forest ($P=0.513$). The extreme gradient boosting tree was the best model, achieving statistically significant improvement over the C5.0 model ($P=0.003$) and the random forest model ($P=0.006$) at an adjusted Bonferroni *P* value at 0.008.

Conclusions: The development of machine learning algorithms significantly improved the accuracy and reliability of current approaches to the evaluation of the linguistic accessibility of Chinese health information, especially Chinese health translations in relation to original health resources. Although the new algorithms developed were based on Chinese health resources, they can be adapted for other languages to advance current research in accessible health translation, communication, and promotion.

(JMIR Med Inform 2021;9(10):e30588) doi:10.2196/30588

KEYWORDS

machine learning; health translation; Chinese health resources

Introduction

Translation serves as an important educational tool for health education and health promotion among multilingual and multicultural populations [1-3]. Health literacy research shows that improving the linguistic accessibility and understandability of health translations can have an important impact on the uptake of health recommendations by medical professionals and health authorities [4]. Current approaches to multicultural health resource evaluation are chiefly qualitative and use clinically developed guidelines or the judgement of health professionals [5,6]. There are several limitations to these approaches. First, there is the potential inconsistency of evaluation among medical professionals. Second, generalized, principled evaluation of health agencies tends to have low adaptability or flexibility, but users of health translations represent vulnerable populations

who vary in their language, culture, education backgrounds, cognitive abilities, and health literacy [7]. Third, evaluation by experts requires longer timeframes, particularly with large quantities of translation resources. In situations of health emergencies which require rapid, more regular communication of information of health risk prevention or management, this can be technically challenging. Finally, logistically, the expert evaluation of health translations in minority languages can be costly [8,9] or is simply not available when there is a lack of suitably qualified medical professionals with adequate knowledge and understanding of minority languages as the target languages of health translations.

In medical and health research, an established approach to the quantitative evaluation of health education resources is the use of readability tools. Some widely used comparable readability tools include Flesch reading ease score, the Gunning fog index, Flesch-Kincaid grade level readability, the Coleman-Liau index, the SMOG index, the automated readability index, and the Linsear Write formula [10-14]. The mathematical design and the functions of these readability tools primarily focus on 3 large dimensions of the linguistic accessibility of health resources for the public: morphological complexity (average number of syllables per word, average characters per word, and average number of letters per word), syntactic complexity (average sentence length and average number of words in sentences), and semantic complexity (percentage of hard words). These readability tools offer fast, convenient measurements of original health resources and provide instant evaluation of the suitability of a new health text for readers of a certain education level. They can also be applied in the study of translated resources in English or European languages.

The limitation of these readability diagnostic tools is also known. First, the measurement of linguistic or textual complexity at the morphological level is based on the calculation of syllables or letters which can hardly be applied in languages that use different alphabets or symbols. For example, East Asian languages like Chinese, Japanese Kanji, and Korean use strokes instead of letters or syllables. Even within the same language, written language varieties; for example, the traditional or simplified versions of Chinese, or the hiragana, katakana, and romanji of Japanese can compound the measurement of linguistic complexity significantly. Second, existing medical readability formulae tend to exploit the orthographical or sentential structures of texts. The design of these readability formulae assumes that linguistic readability can be explained or controlled by reducing the length of individual words, sentences, or the frequency of occurrence of specialized terminology. This assumption has simplified the complexity of the cognitive mechanism that underlies the reading and understanding of specialist health information [15,16]. This includes notably textual logic or coherence devices, such as pronouns, personal pronouns, or conjunctions. In corpus translation studies, the enhanced use of these functional linguistic devices is known as translational features or translationese [17-20]. However, whether these functional categories in translationese can be deployed systematically in professional health translations to increase the readability of translated health resources remains underexplored.

Our study developed new techniques to improve current approaches to Chinese health translation readability evaluation. First, we increased the dimensions of quantitative analyses by incorporating functional categories including pronouns, personal pronouns, conjunctions, and negative conjunctions to the existing measurements of morphological and sentential structural complexity. Second, we adapted the measurements of morphological complexity of European languages to character-based Asian languages. This included adding new morphological measurements of Chinese 2-character words, Chinese 3-character words, average strokes per character, high-stroke characters, low-stroke characters, and middle-stroke characters. Chinese 2- and 3-character words represent the common Chinese vocabulary; meanwhile, 4-character words are more associated with idioms or idiomatic expressions, and words of 4 or more characters are likely to be either specialized terminology, proper nouns, or translated expressions. The morphological complexity of Chinese characters is measured by the number of strokes in each character. High-stroke characters are comparable to polysyllabic words in English or European languages.

In our study, a third adaption to the existing medical readability tools involved the expansion of semantic categories. Among 7 widely used medical calculators, only the Linsear Write formula and the Gunning fog index incorporate “hard,” “difficult,” and “easy” words in the calculation of the linguistic readability of health translations. The addition of words of varying cognitive difficulty represents an advance from the quantification of health information readability based on word or sentence length, such as is done in the Flesch reading ease score, Flesch-Kincaid grade level readability, Coleman-Liau index, SMOG index, and automated readability index. However, the addition of easy versus hard words in Linsear Write formula and Gunning fog has an inherent methodological limitation, notably the lack of a clear, consistent definition of easy or hard words. The interpretation of lexical difficulty is open to the understanding of the users of these readability tools, thus causing inconsistency in the evaluation results among users of the same tool or between different readability tools. In this study, this inherent variability of the Linsear Write formula and Gunning fog was controlled by 6 clearly defined, quantifiable linguistic features based on cognitive and corpus linguistic research on their relevance for the semantic complexity of texts.

Table 1 lists these semantic categories: type and token ratio, density of content words, difficulty words, ratios of noun phrases, normalized frequency of noun phrases, and sentences with complex semantic categories (ie, polysemes). Here, the definition of difficult words is based on the extraction of the 3000 most common Chinese character words in the Academia Sinica Balanced Corpus of Modern Chinese developed by Academia Sinica. Character words which are not listed in the top 3000 words are retrieved as difficult Chinese words. Although the current threshold level of easy versus difficult words based on the first 3000 words in the balanced Chinese corpus is subjective, this corpus approach provided a more transparent and consistent reference point for different users of the new readability system (Table 1). The use of standardized semantic categories instead of absolute values as in the Linsear

Write formula and Gunning fog can help reduce the impact of the length of the texts on the readability evaluation results. This is particularly useful for the evaluation of the Chinese health translations collected in the World Health Organization (WHO) website, which are of varying lengths based on the health topics and genres. The standardized semantic categories are type-token ration (TTR; proportion of different words within the total words of a text), density of content words (the proportion of content words within the total words of a text), ratios of noun phrases (proportion of noun phrases within the total words of a text), and normalized frequency of noun phrases (proportion of noun phrases per 10,000 words).

Table 1. New multidimensional framework of the linguistic readability of health translations in Chinese.

Category	Features
Morphology	Two-character words, three-character words, average strokes per character, high-stroke characters (above 21 strokes), low-stroke characters, middle-stroke characters (11-20 strokes)
Sentences structure	Average sentences per paragraph, average words per sentence, simple sentences
Semantics	Type-token ratio, content words, density of content words, difficult words (beyond the most common 3000 words), ratios of noun phrases, normalized frequency of noun phrases, sentences with complex semantic categories (polysemes)
Logic and coherence	Conjunctions, positive conjunctions, negative conjunctions, personal pronouns, pronouns, adverbs of negation

Methods

Data Collection

With an increasing number of health translations accessible on the internet and rapidly developing computational techniques, the development of cost-effective, robust algorithms for the computerized evaluation of the linguistic accessibility of health translations has become possible. This study explored machine learning techniques to effectively analyze and diagnose the linguistic accessibility of health translations by professional translators of the WHO. Two comparable corpora were constructed containing professional Chinese health translations developed by the WHO (350 full-length translations). The reference materials used to compare with Chinese health translations were original, public-oriented health educational resources published by China Center for Disease Control and Prevention (CCDC). These resources are regarded as authoritative health information widely disseminated by governmental organizations, industrial sectors, and the media. The use of original Chinese health education resources instead of human evaluation of the linguistic accessibility of health resources has both its methodological advantages and limitations. First, human evaluation is known to be susceptible to inconsistency unless there are clear, well-defined criteria of selecting human evaluators, such as age, gender, educational background, health literacy, and cognitive abilities. Although this could help limit the variability in the human evaluation, the evaluation results can be hardly representative of larger, more diverse populations. Another practical issue is the lack of well-established, national guidelines of the level of health educational resources for the public in China. This stands in contrast with English-speaking countries where health authorities provide clinical guidelines or recommendations for the suitable readability level of health resources to ensure access to health information by the greater population. Furthermore, few Chinese health resources have been assessed by international health website accreditation authorities (HON.net). This made it difficult to identify and collect health education resources in Chinese that could meet international health resource development guidelines or clinical and research-based recommendations. The use of CCDC resources was based on 2 considerations. The first was the authority of these resources in China, as the CCDC is the national disease prevention authority and the health educational materials by the CCDC have wide circulation within the country. Second, the quality control provided by the CDCC website content editor ensures the usability and understandability of the resources for the public in China. In this study, during the construction of the subcorpus of original Chinese health educational resources, native Chinese speakers were instructed to select and collect resources intended by the public, rather than technical materials written for medical health professionals, such as disease epistemology or clinical research. This was facilitated by the design of the CCDC website, which has designated sections for public health education resources to describe and explain complex or common diseases and symptoms. This data collection strategy has its limitation: there is a lack of national guidelines of health resource development, and thus the readability or accessibility of health content is not regulated or controlled by national or organizational standards. The content difficulty of these original Chinese health resources on the CCDC website may well be mixed. To overcome this issue, a large number of full-length original Chinese articles were randomly collected from the CCDC website to match the corpus of Chinese health translations of the WHO. Descriptive statistics of original and translated Chinese health texts are given in Multimedia Appendix 1.

Statistical Analysis

Multimedia Appendix 1 shows the differences between the 2 comparable corpora of translated and original health resources covering diverse health topics. The Chinese translations were collected from the website of the WHO, and the original Chinese health resources were published on the website of the CCDC. The *P* values were derived from the Mann-Whitney test with SPSS version 20 (IBM Corporation). The results show that there were statistically significant differences between the translated and original Chinese resources in 20 of the total 22 linguistic and textual features studied. It was found that at the morphological level, there were more low-stroke characters in the original Chinese resources (mean 407.76), which was 1.4

times more than in the translated Chinese health resources (mean 285.06). However, there were also more middle- and high-stroke characters (around 1.6 times) in the original Chinese health resources than in the translated Chinese ones. As a result, the average stroke per character of the original Chinese (mean 7.86) was significantly higher than that of the translated Chinese materials (mean 7.71; *P*=.01). There were more 2- and 3-character words in the original Chinese health resources (mean 2-character words 160.73; mean 3-character words 13.77) than in translated ones (mean 2-character words 114.49; mean 3-character words 8.29). Most of the modern Chinese lexis is made of 2 or 3 characters, suggesting that the lexical familiarity of original Chinese resources could be higher than the translation materials. Second, in terms of information load, the average TTR of the original Chinese texts (mean 0.59) was significantly lower than that of the translated texts (mean 0.62; *P*<.001). Similarly, the average words per sentences of the original Chinese texts (mean 10.83) was significantly lower than that of the translated Chinese resources (mean: 11.9: *P*<.001).

By contrast, the average sentences per paragraph of the original Chinese texts was almost double that of the Chinese health translations (mean 3.32; *P*<.001). This suggests that the translated health materials were longer and contained more information than did the original Chinese health texts and that the original Chinese materials featured longer paragraphs with more sentences despite the average sentence lengths being shorter. Another interesting finding was that there were more single sentences in the translated Chinese materials (mean 0.46) than in the original Chinese health texts (mean 0.32). This may be explained by the more frequent use of logical and coherence words in the original Chinese health texts than in the WHO translations. For example, there was a statistically significant higher use of conjunctions in the original Chinese (mean 14) than in the Chinese health translations (mean 11.53; *P*=.01), there were more negative conjunctions in the original Chinese (mean 2.03) than in the translations (mean: 1.44, *P*=.03), and there were more pronouns and personal pronouns in the original Chinese (mean: 2.36 for pronouns; 1.12 for personal pronouns) than in the Chinese health translations (pronouns: mean 1.47, *P*=0.01; personal pronouns: mean 0.7, *P*=.04). At the semantic level, although the difference between the original and translated Chinese health resources was not significant in terms of the ratio of noun phrases, the normalized frequency of noun phrases was much higher in the original (mean 321.89) than in the translated Chinese health texts (mean 314.58; *P*=.04). It was useful to notice that the mean of sentences with complex semantic categories (polysemes) was statistically higher—almost double—in the original Chinese resources (mean 14.42) compared to the translated Chinese ones (mean 7.61; *P*<.001). Finally, the density of content words in the original Chinese (mean 0.83) was statistically higher than that of the translated health texts (mean 0.81).

The statistical comparison (using the nonparametric test for 2 independent samples: Mann-Whitney test) indicated a mixed feature of the linguistic accessibility of original and translated Chinese health resources. It suggested that at the morphological level, the original Chinese health texts are more complex than are translated health texts, as the average number of strokes in characters was higher in the original Chinese texts than in the translated Chinese ones. However, this issue was somewhat offset by the higher use of more familiar lexis in the original text compared to the translated health materials, as the mean of 2- and 3-character words was significantly higher in the original Chinese health texts than in the Chinese health translations. Next, at the semantic level, the statistical comparison showed that the original Chinese health texts were more complex than the Chinese health translations, as evidenced by the higher density of content words and the higher mean of sentences with polysemes. However, this issue occurred in conjunction with the higher information load of Chinese health translations than of the original health texts, as illustrated by the higher TTRs and higher average words per sentences of health translations.

In terms of the logical structure and coherence of sentences, although there were more simple sentences in the Chinese health translations and more compound sentences in the original Chinese health texts, the latter featured more coherence devices such as pronouns and personal pronouns to assist with the reading and understanding of the original Chinese health texts. This mixed outcome of the comparison between the 2 comparable corpora suggests that the assessment of the linguistic accessibility of health resources should be balanced and multidimensional to avoid any partial, biased assessment outcome. The proposed multidimensional corpus analysis contrasts with the use of medical readability tools which focus on the morphological and syntactic features of health texts, such as word length in letters or syllables or sentence lengths in words. The machine learning modeling results to be shown demonstrate that semantic, logical, and coherence features are also equally important for building effective evaluation models of the linguistic accessibility of Chinese health translations.

Development of Machine Learning Models

Machine learning models are different from conventional statistical methods. First, machine learning does not require a normal distribution to fit the data to existing statistical models. Machine learning is essentially data driven and is free to learn any functions underlying the training data without any implicit assumptions. This is especially the case for tree-based machine learning algorithms, such as gradient boost trees, decision trees, and random forest. With conventional statistical methods, the values of the independent variables and dependent variables must be both present to model their relationships, whereas with machine learning, once the algorithm has been well tested and validated, it can be used to predict the outcome of the target variable based on the information collected from the predictor variables. That is, machine learning can be used to make high-precision predictions. This suits the purpose of this study, whose aim was to develop a cost-effective health translation accessibility prediction model which can predict whether a certain health translation differs significantly from the original Chinese health education resources developed for the public of native Chinese speakers, for example, the health education resources we collected carefully from the website of the CCDC as the national health authority in China.

Effective machine learning models can instantly detect any potential reading barriers caused by linguistic accessibility issues

of health translations, allowing translation and international health agencies to revise translations before their release to the public. Linguistically more accessible health translations in turn can help achieve better outcomes of health education programs and campaigns among vulnerable populations with limited English proficiency and low education and health literacy levels. In this study, random forest and extreme gradient boosting tree (XGBoost tree) were used to solve a classification problem: that is, to predict whether a certain Chinese health text is more likely to be of the linguistic accessibility level of an original Chinese health text or of a Chinese health translation based on the modelling of the linguistic features as the predictor variables to be extracted from an unlabeled health text in Chinese.

To overcome model overfitting, 5-fold cross-validation on the whole data set was conducted. In each 5-fold cross-validation, the entire data set was divided into 5 portions of approximately equal numbers. Four folds of data were used as the training data to develop the machine learning models, and the remaining fold was used as the test data set to calculate the model performance metrics including area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy. After the iterations, each fold was used as the test set exactly once. The 5-fold cross-validation can help detect overfitting machine learning models that have large differences in the performance metrics of the 5 test data sets.

In this study, hyperparameter tuning of XGBoost tree involved the following steps. The maximum tree depth for base learners (max_depth) controls the depth of the tree. The greater the depth is, the more complex is the model and the higher are chances of the model overfitting. There is no standard value for adepts. Larger data sets require deep trees to learn the rules from a complex data set. The value ranges between 0 and infinity. In the cross-validation process, we set max_depth to the default value of 6. The number of estimators or boosted trees was set to the default value of 100. The minimum sum of instance weight needed in a child node (min_child_weight) is another effective overfitting prevention method. It is calculated by second-order partial derivatives and ranges between 0 and infinity. The larger the value, the more conservative the algorithm is. This was set to the default value of 1 in this study. The maximum delta step (max_delta_step) specifies the maximum step size that a leaf node can take. It ranges between 0 and infinity, and increasing the positive value will make the update step more conservative. It was set to 1 in this study. The learning objective was set to binary logistic regression, as the target variable had 2 outcome categories: the original Chinese health education resources and the translated Chinese health resources. The hyperparameter subsample refers to the subsample ratio of the training instance. For example, setting the subsample to 0.5 means that the algorithm randomly collects half of the entire data set to build the tree model. This method can prevent overfitting. The value of the subsample was set to 0.5 from its typical range of 0.5 to 0.8. Eta refers to the machine learning rate at which the algorithm learns the latent patterns and structures in the training data set. Smaller eta leads to slower computation and thus prevents overfitting. Smaller etas can be compensated for by increasing the number of boosted trees or estimators. Typically, the value of eta lies between 0.01 and 0.3, and 0.1 was set as the default value in this study. The hyperparameter colsample_bytree controls the number of features or variables supplied to a tree model, with a typical value ranging between 0.5 to 0.9, and it was set to 0.5 in this study. Lastly, α and λ values which control L1 and L2 regularization, were set to 1 and 0, respectively, to prevent overfitting.

Similar to XGBoost tree, random forest is another powerful ensemble learning technique that outperforms single learning algorithms in machine learning model development. In random forest, decision trees are used as the base learner, and bootstrapping aggregation combines these decision trees together to achieve high prediction accuracy. The minimum number of samples and training data required to be at a leaf node (min_samples_leaf) was set to 3. The maximum depth was set to 6. The number of features to use for splitting was set to auto. In the model construction process, the ensemble learning methods selected to increase the prediction accuracy included bootstrapping, bagging, and extremely randomized trees. In the process of hyperparameter optimization, on each iteration, the algorithm chooses a difference combination of the features. The maximum number of iterations was set to 1000, and the maximum evaluations was set to 300.

Results

In the evaluation of the performance of the new machine learning models developed using XGBoost, random forest, and C5.0 decision tree, logistic regression was used as the baseline, as logistic regression has been used widely in both conventional statistical methods and traditional machine learning modelling. Table 2 shows that the logistic regression model was statistically significant. Table 3 shows the entered selection of important predictor variables in the final logistic regression model. It was found that among the initial 22 predictor variables (Multimedia Appendix 1), 4 predictor variables were identified as large contributing variables to the logistic regression model. These were average sentences per paragraph, middle-stroke characters, difficult words, and conjunctions. When the original Chinese health resources were used as the reference category, the Exp (B) values showed that textual features, including as higher average sentences per paragraph, higher middle-stroke characters, and higher use of difficult words were important features associated with the original Chinese health resources. For example, the average sentence per paragraph (P<.001) had an Exp (B) value of 0.44. This means that with the increase of one unit in average sentence per paragraph, the odds of the text being a Chinese health translation over the odds of the text being an original Chinese health text were 44%. Similarly, middle-stroke characters (P=0.03) had an Exp (B) value of 0.97. This suggests that with other variables being the same and with the increase of 1 middle-stroke character, the odds of the text being a Chinese health translation over the odds of the text being an original Chinese health text were 97%. Difficult words (P=.04) as the predictor variable had an Exp (B) value of 0.976. This means that with the increase of one difficult word, the odds of the text being a Chinese health translation were 2.4% lower than the odds of the text being an original Chinese health text. This finding suggests that higher use of difficult words is an

important feature of original Chinese health educational resources when compared to Chinese health translations. By contrast, the predictor variable of conjunctions (*P*=.02) had an Exp (B) of 1.122, which means that with the other variables being the same and with the increase of 1 conjunction, the odds of the text being a Chinese health translation were 12.2% higher than the odds of the text being an original Chinese health education text. This corpus statistical finding can be explained by the theoretical hypotheses of translation studies such as translationese: the increased use of linguistic devices like conjunctions enhanced the textual cohesion of translated materials.

Table 2. Variables in the equation for the original reference category.

Variables	B	SE	Wald	Sig.[a] (*P* value)	Exp(B)	95% CI for Exp (B)	
						Lower bound	Upper bound
Intercept	4.311	2.432	3.142	.07			
Average sentences per paragraph	–0.821	0.149	30.503	<.001	0.44	0.329	0.589
Frequency of noun phrases per 10,000 words	–0.008	0.005	2.732	.10	0.992	0.983	1.001
Average words per sentences	0.249	0.162	2.367	.12	1.282	0.934	1.76
Middle-stroke characters	–0.031	0.014	4.518	.03	0.97	0.943	0.998
Content words	0.005	0.009	0.291	0.59	1.005	0.987	1.023
Difficult words	–0.025	0.012	4.084	.04	0.976	0.953	0.999
Two-character words	0.012	0.011	1.053	.31	1.012	0.989	1.035
Conjunctions	0.115	0.047	5.93	.02	1.122	1.023	1.232
Sentences with complex semantic categories	0.006	0.064	0.008	.93	1.006	0.888	1.14
Ratio of noun phrases	0.719	0.995	0.521	.47	2.052	0.292	14.435
Pronouns	–0.1	0.148	0.461	.50	0.905	0.677	1.209
Personal pronouns	–0.071	0.198	0.127	.72	0.932	0.632	1.375
Negative conjunctions	0.12	0.146	0.677	.41	1.127	0.847	1.5

[a]sig: significance.

Table 3 shows the comparison of the AUC of the 3 machine learning models in comparison with the traditional logistic regression model: C5.0 decision tree has an AUC of 0.969, which is followed by random forest with an AUC of 0.957 and XGBoost tree with an AUC of 0.914. To evaluate the statistical significance between these models in terms of AUC improvement, the pairwise corrected resampled test was conducted as shown in Table 4. To overcome multiple comparison, the significance level was adjusted to 0.008 using Bonferroni correction. The result showed that the 3 new machine learning models (ie, C5.0 decision tree, random forest, and XGBoost tree) significantly improved the performance of the prediction, as their AUCs were significantly higher than the AUCs of logistic regression. There was no significant difference between C5.0 decision tree and random forest, which were significantly more precise than was XGBoost tree.

Table 3. Mean area under the receiver operating characteristic curve.

Test result variable(s)	AUC[a]	SE	Asymptotic sig.[b] (*P* value)	Asymptotic 95% CI	
				Lower Bound	Upper Bound
XGBoost tree[c]	0.914	0.019	<.001	0.878	0.951
Random forest	0.957	0.014	<.001	0.93	0.984
C5.0 decision tree	0.969	0.012	<.001	0.945	0.992
Logistic regression (baseline)	0.768	0.025	<.001	0.718	0.818

[a]AUC: area under the curve.

[b]sig: significance.

[c]XGboost tree: extreme gradient boosting tree.

Table 4. Paired-sample *t* test of the area under the receiver operating characteristic curve.

Test result pair(s)	Asymptotic		AUC[a] difference	SE difference[b]	Asymptotic 95% CI	
	z	Sig. (2-tailed *P* value)[c,d]			Lower bound	Upper bound
XGBT[e]-RF[f]	–2.74	.006	–0.043	0.179	–0.073	–0.012
XGBT-LR[g]	6.118	<.001	0.146	0.209	0.099	0.193
XGBT-C5[h]	–2.967	.003	–0.054	0.175	–0.09	–0.018
RF-LR	7.405	<.001	0.189	0.197	0.139	0.239
RF-C5	–0.655	.51	–0.011	0.161	–0.046	0.023
LR-C5	–8.094	<.001	–0.201	0.193	–0.249	–0.152

[a]AUC: area under the curve.

[b]Under the nonparametric assumption.

[c]Null hypothesis: true area difference = 0.

[d]A *P* value <0.008 is statistically significant (using Bonferroni correction). *P* values were derived from the pair-wise corrected resampled *t* test.

[e]XGBT: extreme gradient boosting tree.

[f]RF: random forest.

[g]LR: logistic regression.

[h]C5: C5.0 decision tree.

Discussion

Next, through successive permutation, we examined the impact of textual features as predictor variables on the change in percentage of AUC in the best-performing machine learning algorithms, the C5.0 decision tree, and the baseline algorithm of logistic regression. Table 5 shows that a number of textual features contributed to changes in the mean AUC of logistic regression (0.768), notably, average sentences per paragraph (4.8%), simple sentences (2.6%), normalized frequency of noun phrases per 10,000 words (2.4%), average strokes per character (1.6%), conjunctions (1.5%), content words (1.3%), low-stroke characters (1.2%), and sentences with complex semantic categories (1.1%). By contrast, C5.0 decision tree which had a statistically better AUC (0.914) and featured a different set of textual features as key contributors to changes of the algorithm's AUC. These included density of content words (7.3%), average words per sentence (3.2%), simple sentences (2.9), negative conjunctions (2.1%), positive conjunctions (2%), personal pronouns (1.8%), average sentences per paragraph (1.1%), high-stroke characters (1.1%), and TTR (1.1%). Only 2 textual features were identified as important contributors to both algorithms (causing a 1% or more decrease in AUC): simple sentences and average sentences per paragraph. Both simple sentences and average sentences per paragraph are measurements of the syntactic complexity of sentences. Existing medical readability formulae attempt to capture these features using average sentence length or average number of words in sentences. Morphological complexity was measured with different natural language features in the 2 algorithms. Logistic regression measured morphological complexity using average strokes per character and low-stroke (under 10 strokes) characters. C5.0 decision tree measured morphological complexity using high-stroke (over 20 strokes) characters.

Table 5. Changes in the area under the receiver operator characteristic curve through successive permutation of features.

Feature	LR[a]	C5[b]
Sentences with complex semantic categories	0.011	0
Content words	0.013	0
Low-stroke characters	0.012	0
Adverbs of negation	0.01	0
Average strokes per character	0.016	0
Two-character words	0	0.001
Conjunctions	0.015	0.001
Ratios of noun phrases	0.009	0.001
Normalized frequency of noun phrases per 10,000 words	0.024	0.001
Three-character words	0	0.002
Difficult words	0.01	0.002
Middle-stroke characters	0.001	0.004
Pronouns	0.001	0.008
Average sentences per paragraph	0.048	0.011
Type-token ratios	0.005	0.011
High-stroke characters	0.01	0.011
Personal pronouns	0.01	0.018
Positive conjunctions	0.009	0.02
Negative conjunctions	0.007	0.021
Simple sentences	0.026	0.029
Average words per sentences	0.007	0.032
Density of content words	0.01	0.073

[a]LR: logistic regression.

[b]C5: C5.0 decision tree.

Both machine learning algorithms identified and explored additional linguistic dimensions which were not studied in exiting medical resource readability assessment formulae. The first was information load, which refers to the amount and complexity of information contained in the texts. Logistic regression measured information load using natural language features, such as normalized frequency of noun phrases per 10,000 words and content words. Both categories contributed more than 1% of the changes in the AUC of logistic regression. Noun phrases are phrases which contain nouns and function as nouns in a sentence, and they are used extensively in medical and scientific writing. Higher normalized frequencies of noun phrases can significantly increase the information load of scientific discourse. C5.0 decision tree by contrast measured the information load of translated and original Chinese health resources using the density of content words, which was the percentage of content words of both content and function words in the Chinese texts. Content words include part-of-speech categories, including nouns, adjectives, adverbs, and verbs, whereas functional words comprise auxiliary verbs, pronouns, articles, and prepositions. A higher density of content words was found as a statistically significant feature of original Chinese health resources (mean 0.83) when compared to translated Chinese health resources (mean 0.81). TTR is another widely used measure in corpus linguistics to measure lexical diversity or richness. A higher TTR indicates an increased variety of words, and this was a significant feature of Chinese health translations (mean 0.62) when compared to the original Chinese health resources (mean 0.59).

Another dimension of linguistic features which was identified by the 2 machine learning algorithms based on natural language features was textual coherence and logical structure, which are not included in existing medical readability formulae. Logistic regression measured textual structure by using conjunctions, whereas the C5.0 decision tree algorithm exploited natural language features, such as negative conjunctions, positive conjunctions, pronouns, and personal pronouns. The original Chinese health resources featured a higher use of all these linguistic classes to heighten the logical structure of the Chinese health texts, whereas the translated Chinese health resources exhibited a more conservative use of these functional lexical categories. This finding cannot be explained by the hypothesized translationese or universal translation pattern of lexical simplification or normalization which is achieved through an increased use of functional devices such as conjunctions and pronouns. Rather, this may be a product of the influence from

the source language, as English scientific discourse tends to use more passive sentence structures, whereas pronouns and personal pronouns are more common in everyday Chinese texts that use more direct, positive sentences. These linguistic features are related to the cognitive and logical properties of health texts. Although these textual features have not been incorporated into medical readability formulae, they are highly relevant to widely used health education resource development guidelines, for example, the Patient Education Materials Assessment Tool, which is a systematic method developed by the US Department of Health and Human Services to evaluate the understandability and actionability of patient education materials. Positive conjunctions and adverbs of negation may impact the logical sequence of health information, which is specified in the Patient Education Materials Assessment Tool as a key criterion for assessing the linguistic understandability of health educational resources.

Linguistic accessibility has an important impact on the reception and utilization of translated health resources among multicultural and multilingual populations with a high proportion of immigrants. Linguistic understandability of health translation has been understudied. Automated predictive analyses of the linguistic accessibility of new health translations before their release to the public can significantly improve the cost-effectiveness and efficiency of bilingual, multilingual health education programs and the use of health translation resources by the public. This paper introduced machine learning techniques to the study of the linguistic accessibility of health translations by comparing Chinese translations of the WHO materials with the original Chinese health resources developed by the Chinese health authorities. Three new machine learning models (XGBoost tree, random forest, and C5.0 decision tree) were developed and compared in terms of their accuracy, AUC, sensitivity, and specificity with the traditional logistic regression modeling being used as the baseline. The selection of textual features was based on existing research on corpus-based translation studies, cognitive linguistics, health literacy, and public health education. A number of textual and linguistic features were selected, which included morphological features, such as 2- and 3-character words, average strokes per character, low-stroke characters, and middle-stroke or high-stroke characters; features of sentential structures, such as average sentences per paragraph, average words per sentences, sentences with complex semantic categories, and single sentences; semantic features, such as TTRs, content words, density of content words, difficult words, and normalized frequency of noun phrases; and textual logic and coherence features, such as conjunctions, negative conjunctions, personal pronouns, and pronouns. Five-fold cross-validation was conducted in the model development process to ensure the reliability and replicability of the new machine learning models.

In the evaluation of the performance of the machine learning models, cross-model comparison was conducted. To counteract the issue of multiple comparisons and the risk of erroneous inferences, the significance level of the paired model comparison was adjusted to 0.008 using Bonferroni correction. The pairwise corrected resampled tests showed that C5.0 decision tree, random forest, and XGBoost tree all outperformed logistic regression with statistically higher AUCs. The impact of linguistic features on the AUC of the best-performing model, C5.0 decision tree, and the baseline logistic regression model was analyzed through eliminating 1 textual feature at a time from the whole set of textual features built within each algorithm. It was found that, like most existing linguistic readability evaluation formulae developed in medical research, both C5.0 decision tree and the baseline logistic regression model measured the morphological and syntactic complexity of health texts. Medical readability formulae focus on the average number of letters or syllables within words, or the average number of words within sentences, while machine learning models measure morphological and syntactic complexity using natural language features. Furthermore, machine learning models increase the dimensions of the analysis of linguistic accessibility of health texts.

The 2 new dimensions exploited by machine learning were information load and textual coherence. Information load was measured by natural language features including normalized frequency of noun phrases, density of content words, and TTR. Textual coherence and logical organization were measured by the 2 large categories of functional words, conjunctions (positive, negative) and pronouns (including personal pronouns). These quantitative models can serve as highly accurate, automated analytical tools to help predict linguistic accessibility of health translations in Chinese and represent a methodological advance from existing qualitative approaches in terms of the reliability, efficiency, and cost-effectiveness of the evaluation. The development of machine learning algorithms significantly improves upon the accuracy and reliability of current approaches to the evaluation of the linguistic accessibility of Chinese health information, especially Chinese health translations in relation to original health resources. Although the new algorithms developed were based on Chinese health resources, they can be adapted for other languages to advance current research in accessible health translation, communication, and promotion.

Automated predictive analyses of the linguistic accessibility of new health translations before their release to the public can significantly improve the cost-effectiveness, efficiency of bilingual and multilingual health education programs, and the use of health translation resources by the public. We developed new machine learning algorithms to help predict linguistic accessibility of health translations in Chinese, which represents a methodological advance from existing qualitative approaches in terms of the reliability, efficiency, and cost-effectiveness of the evaluation.

Acknowledgments

This paper is supported by The University of Sydney and University of Geneva Global Research Partnership Award.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mann-Whitney test of original and translated Chinese health information.
[DOCX File , 18 KB - medinform_v9i10e30588_app1.docx]

References

1. Regmi K, Naidoo J, Pilkington P. Understanding the processes of translation and transliteration in qualitative research. International Journal of Qualitative Methods 2010 Mar 01;9(1):16-26 [FREE Full text] [doi: 10.1177/160940691000900103]
2. Tsai J, Choe J, Lim J, Acorda E, Chan N, Taylor V, et al. Developing culturally competent health knowledge: issues of data analysis of cross-cultural, cross-language qualitative research. International Journal of Qualitative Methods 2016 Nov 29;3(4):16-27 [FREE Full text] [doi: 10.1177/160940690400300402]
3. Ho S, Holloway A, Stenhouse R. Analytic methods' considerations for the translation of sensitive qualitative data from Mandarin into English. International Journal of Qualitative Methods 2019 Aug 09;18:160940691986835 [FREE Full text] [doi: 10.1177/1609406919868354]
4. Anne B, Mira K. Community accessibility of health information and the consequent impact for translation into community languages. Translation and Interpreting 2011;3(1):58-75 [FREE Full text]
5. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information. Patient Education and Counseling 2014 Sep;96(3):395-403 [FREE Full text] [doi: 10.1016/j.pec.2014.05.027]
6. Lipari M, Berlie H, Saleh Y, Hang P, Moser L. Understandability, actionability, and readability of online patient education materials about diabetes mellitus. Am J Health Syst Pharm 2019 Jan 25;76(3):182-186. [doi: 10.1093/ajhp/zxy021] [Medline: 31408087]
7. Ji M, Liu Y, Zhao M, Lyu Z, Zhang B, Luo X, et al. Use of machine learning algorithms to predict the understandability of health education materials: development and evaluation study. JMIR Med Inform 2021 May 06;9(5):e28413 [FREE Full text] [doi: 10.2196/28413] [Medline: 33955834]
8. Kirchhoff K, Turner AM, Axelrod A, Saavedra F. Application of statistical machine translation to public health information: a feasibility study. J Am Med Inform Assoc 2011;18(4):473-478 [FREE Full text] [doi: 10.1136/amiajnl-2011-000176] [Medline: 21498805]
9. Turner AM, Dew KN, Desai L, Martin N, Kirchhoff K. Machine translation of public health materials from English to Chinese: a feasibility study. JMIR Public Health Surveill 2015;1(2):e17 [FREE Full text] [doi: 10.2196/publichealth.4779] [Medline: 27227135]
10. Flesch R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233. [doi: 10.1037/h0057532] [Medline: 18867058]
11. Gunning R. Readability yardsticks. In: The Technique of Clear Writing. New York: McGraw-Hill; 1968.
12. Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology 1975;60(2):283-284. [doi: 10.1037/h0076540]
13. McLaughlin GH. SMOG grading-a new readability formula. Journal of Reading 1969:639-646.
14. Senter R, Smith E. Automated readability index. Defense Technical Information Center. URL: https://apps.dtic.mil/sti/citations/AD0667273 [accessed 2021-03-20]
15. Kevin H. Investigating Adolescent Health Communication: A Corpus Linguistics Approach. London: Bloomsbury; 2013.
16. Rubin D. Applied linguistics as a resource for understanding and advancing health literacy. In: The Routledge Handbook of Language and Health. London: Routledge; 2014:153-167.
17. Volansky V, Ordan N, Wintner S. On the features of translationese. Digital Scholarship in the Humanities 2013 Jul 03;30(1):98-118 [FREE Full text] [doi: 10.1093/llc/fqt031]
18. Ilisei I. Identification of translationese: a machine learning approach. 2010 Presented at: International Conference on Intelligent Text Processing and Computational Linguistics; March 21-27, 2010; Iaşi, Romania p. A URL: https://doi.org/10.1007/978-3-642-12116-6_43 [doi: 10.1007/978-3-642-12116-6_43]
19. Baroni M. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. Literary and Linguistic Computing 2005 Aug 05;21(3):259-274 [FREE Full text] [doi: 10.1093/llc/fqi039]
20. Redelinghuys K, Kruger H. Using the features of translated language to investigate translation expertise. IJCL 2015 Aug 28;20(3):293-325. [doi: 10.1075/ijcl.20.3.02red]

Abbreviations

CCDC: China Center for Disease Control and Prevention
TTR: type-token ratio

WHO: World Health Organization
XGBoost tree: extreme gradient boosting tree

Edited by C Lovis; submitted 21.05.21; peer-reviewed by M Oakes, S Nagavally; comments to author 17.06.21; revised version received 21.06.21; accepted 02.07.21; published 07.10.21.

Please cite as:
Ji M, Bouillon P
Predicting the Linguistic Accessibility of Chinese Health Translations: Machine Learning Algorithm Development
JMIR Med Inform 2021;9(10):e30588
URL: https://medinform.jmir.org/2021/10/e30588
doi:10.2196/30588
PMID:34617914

Original Paper

Predictability of Mortality in Patients With Myocardial Injury After Noncardiac Surgery Based on Perioperative Factors via Machine Learning: Retrospective Study

Seo Jeong Shin[1*], MS; Jungchan Park[1,2*], MD; Seung-Hwa Lee[3,4], MD; Kwangmo Yang[1,5*], MD; Rae Woong Park[1,6], MD, PhD

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea

[2]Department of Anesthesiology and Pain Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3]Rehabilitation & Prevention Center, Heart Vascular Stroke Institute, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[4]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul, Republic of Korea

[5]Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[6]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Rae Woong Park, MD, PhD
Department of Biomedical Sciences
Ajou University Graduate School of Medicine
206, World cup-ro, Yeongtong-gu
Suwon, 16499
Republic of Korea
Phone: 82 0312194471
Email: veritas@ajou.ac.kr

Abstract

Background: Myocardial injury after noncardiac surgery (MINS) is associated with increased postoperative mortality, but the relevant perioperative factors that contribute to the mortality of patients with MINS have not been fully evaluated.

Objective: To establish a comprehensive body of knowledge relating to patients with MINS, we researched the best performing predictive model based on machine learning algorithms.

Methods: Using clinical data from 7629 patients with MINS from the clinical data warehouse, we evaluated 8 machine learning algorithms for accuracy, precision, recall, F1 score, area under the receiver operating characteristic (AUROC) curve, and area under the precision-recall curve to investigate the best model for predicting mortality. Feature importance and Shapley Additive Explanations values were analyzed to explain the role of each clinical factor in patients with MINS.

Results: Extreme gradient boosting outperformed the other models. The model showed an AUROC of 0.923 (95% CI 0.916-0.930). The AUROC of the model did not decrease in the test data set (0.894, 95% CI 0.86-0.922; *P*=.06). Antiplatelet drugs prescription, elevated C-reactive protein level, and beta blocker prescription were associated with reduced 30-day mortality.

Conclusions: Predicting the mortality of patients with MINS was shown to be feasible using machine learning. By analyzing the impact of predictors, markers that should be cautiously monitored by clinicians may be identified.

(JMIR Med Inform 2021;9(10):e32771) doi:10.2196/32771

KEYWORDS

myocardial injury after noncardiac surgery; high-sensitivity cardiac troponin; machine learning; extreme gradient boosting

Introduction

Myocardial injury after noncardiac surgery (MINS) is associated with cardiovascular events and fivefold increased postoperative mortality, affecting up to the first 2 years after surgery [1]. Recently, MINS is accepted as the leading cause of postoperative mortality [2,3]. Along with the increased risk of mortality, the prevalence is also high, reported to be above 20%

[2,3]. Many previous studies have reported risk factors for the occurrence of MINS [4-7], but relatively less attention has been given to perioperative factors that are associated with mortality in patients who were diagnosed with MINS. We reported perioperative factors that affect mortality after MINS [8-11]. However, our previous studies evaluated variables independently and not in a comprehensive manner.

In this study, we trained and evaluated machine learning models by leveraging the risk factors of patients with MINS and aimed to find a model with the best performance. Furthermore, we validated the performance of the model with the test data set that was curated by the same method with the training data set. By quantifying and comparing the effect of each variable on the predictive performance of the model, we developed a mobile app that predicts mortality in patients with MINS. Our findings may benefit a comprehensive understanding of patient characteristics related to mortality in patients with MINS.

Methods

The Institutional Review Board at Samsung Medical Center forwent the approval for this study and the necessity to obtain informed consent for access to the Samsung Medical Center Troponin in Noncardiac Operation (SMC-TINCO) registry (SMC 2019-08-048) and the test data set for validation (SMC 2021-03-187), considering that both data sets were curated in deidentified form.

Study Population and Data Curation

Samsung Medical Center is a tertiary referral center with nearly 2000 beds and more than 49,000 cases of surgeries performed every year. Additionally, they provide the clinical data warehouse called “Darwin-C,” which allows any researcher in the institution to automatically extract the deidentified data from this electronic medical record archive system (Multimedia Appendix 1). Using the “Darwin-C” system, we generated the SMC-TINCO registry (KCT0004244) and used it in this study. The SMC-TINCO contains consecutive data of 43,019 patients who had at least one inspection of cTn-I before or within 30 days after noncardiac surgery from January 2010 to June 2019.

The medical history was summarized by reviewing the preoperative assessment sheet, and the names and meanings of 44 features in the data sets are listed in Multimedia Appendix 2. The death state of the clinical data warehouse is consistently validated and updated from the National Population Registry of the Korea National Statistical Office.

The routine cTn-I assay of SMC was institutionally updated to high-sensitivity cTn-T from July 2019. Based on this change, we generated a data set for testing the model. The data set consists of 6246 adult patients who had postoperative high-sensitivity cTn-T measured within 30 days after noncardiac surgery between July 2019 and January 2021.

Definitions and Study End Points

MINS was defined as peak postoperative cTn elevation above the 99th percentile of the normal limit within 30 days after surgery, but those with evidence of nonischemic etiology such as sepsis, pulmonary embolus, atrial fibrillation, cardioversion, or chronic elevation were not regarded as MINS based on the recent diagnostic criteria [12]. High-risk surgery was identified based on the 2014 European Society of Cardiology/Anesthesiology guidelines [13].

The primary end point was the predictability of 30-day mortality of patients with MINS based on perioperative factors. For the secondary outcome, we also evaluated the predictability of 1-year mortality.

Perioperative Management and cTn Measurements

According to the institutional guidelines, postoperative cTn measurement is not an institutional routine practice. It is performed selectively on patients with one or more of the following major cardiovascular risk factors: heart failure, history of ischemic heart disease, stroke including transient ischemic attack, chronic kidney disease, diabetes mellitus on insulin therapy, or high-risk surgery, but symptoms may be determined at the discretion of the clinician [13].

An immunoassay (Advia Centaur XP, Siemens Healthcare Diagnostics, Erlangen, Germany) with high sensitivity was used for cTn-I. The lower detection limit was 6 ng/L, and 40 ng/L of the 99th percentile was the reference upper limit, as reported by the manufacturer [14]. In the test data set, a high-sensitivity assay of cTn-T (Elecsys, Roche, Basel, Switzerland) was analyzed using cobas e801 (Roche). The 99th percentile reference upper limit for hs-cTn-T was 14 ng/L.

Development of Prediction Models

To compare the performance of prediction models, we investigated the eight widely used machine learning algorithms: extreme gradient boosting (XGB), generalized boosted regression model (GBM), random forests (RF), support vector machines (SVM), classification and regression trees (CART), linear discriminant analysis (LDA), lasso/ridge/elastic net (GLMNET), and k-nearest neighbors (kNN). The hyperparameters of each model were optimized based on a grid search using the area under the receiver operating characteristic (AUROC). Fivefold cross-validation was used in the model development. We evaluated each model according to the accuracy, precision, recall, F1 score, AUROC, and area under the precision and recall curve (AUPRC) values (Multimedia Appendix 3). We validated the performance of the trained model using a new test data set.

Feature importance and Shapley Additive Explanations (SHAP) values were used to present the impact of each feature on the performance of the prediction model. SHA*P* values show the characteristic of deriving a marginal distribution and weighted average by fixing all variables except one and randomly predicting that one to determine its importance [15]. Features are sorted in descending order by which the model contributes to classifying the data. Each patient was represented by one dot on each variable line. The horizontal location of each dot indicated whether the effect of a variable was associated with a higher or lower probability of death. The area on the right indicates the point where SHA*P* value is greater than zero. Variable-specific SHA*P* values >0 indicate an increased risk of death.

Statistical Analysis

Differences were compared using *t* tests and presented as means and SDs in two-group comparisons. Categorical features were presented as numbers with percentages and compared using chi-square or Fisher exact tests. Statistical analyses were performed using R 3.6.3 (R Foundation for Statistical Computing). All tests were two-tailed, and $P<.05$ was considered to indicate statistical significance.

Results

Patient Characteristics

In accordance with the definition of MINS, patients younger than 18 years were excluded from the data sets. Patients who did not have troponin measured after surgery or had abnormal levels and nonischemic etiology, such as chest compression, were also excluded (Figure 1). The baseline characteristics of the study patients with MINS are presented in Table 1. The age and gender of the patients in the training and test data sets showed a similar distribution (Multimedia Appendix 4), but the distribution of surgical types was slightly different. The number of patients in gynecology and urology in the test data set was increased, and other surgeries such as donor transplantation and bronchial dilation also varied (Multimedia Appendix 5). The type of surgery performed on patients in each data set and their mortality are presented in Multimedia Appendix 6.

Figure 1. A flowchart of our retrospective study design. peri-op hs cTn I: perioperative high-sensitivity cTn-I; peri-op hs cTn T: perioperative high-sensitivity cTn-T; Pts: patients; SMN-TINCO: Samsung Medical Center Troponin in Noncardiac Operation.

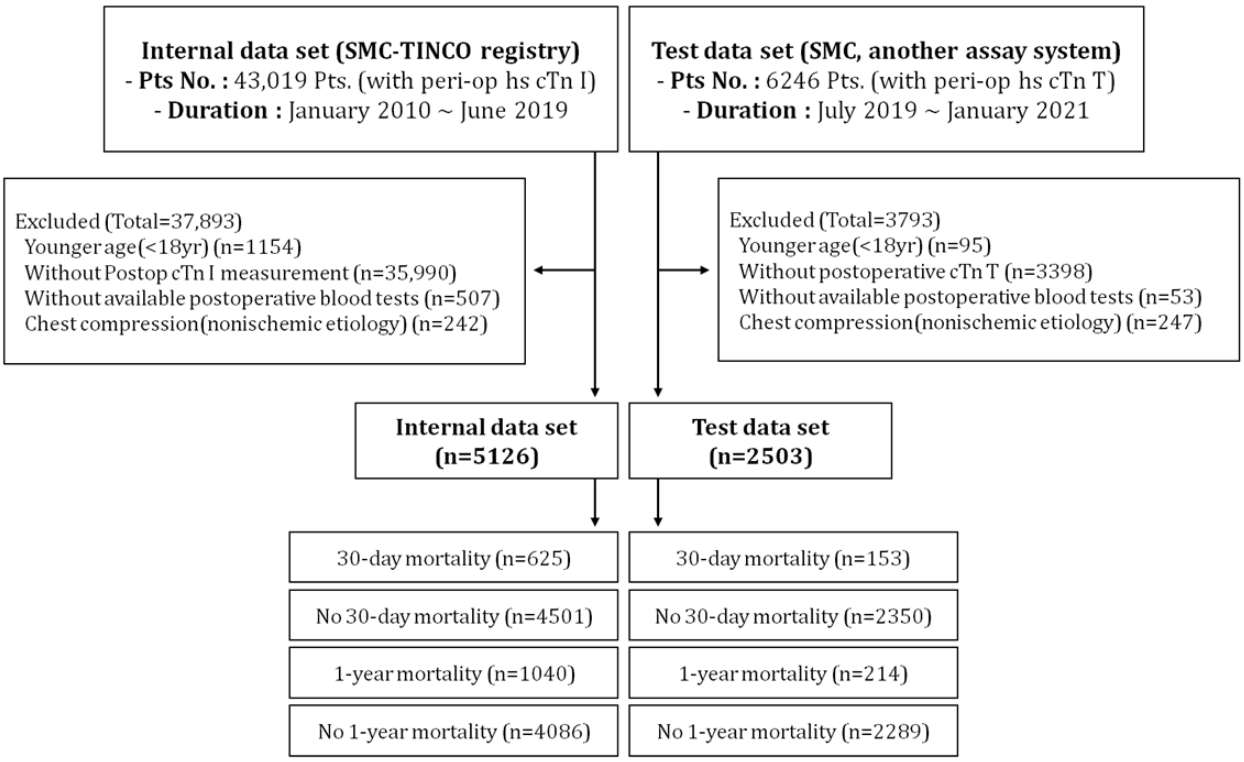


Table 1. Baseline characteristics of patients with myocardial injury after noncardiac surgery according to 30-day mortality.

	Training data set			Test data set		
	No 30-day mortality (n=4501)	30-day mortality (n=625)	*P* value	No 30-day mortality (n=2350)	30-day mortality (n=153)	*P* value
Peak cardiac troponin level (ng/L), mean (SD)	2.3 (19.7)	7.2 (31.7)	<.001	0.1 (0.7)	0.6 (3.4)	.10
Male, n (%)	2673 (59.4)	394 (63.0)	.09	1526 (64.9)	89 (58.2)	.12
Age (years), mean (SD)	65.7 (13.8)	63.1 (14.2)	<.001	68.2 (12.8)	63.3 (13.6)	<.001
BMI, mean (SD)	23.7 (3.8)	22.9 (3.5)	<.001	24.0 (3.8)	22.7 (3.9)	<.001
Diabetes, n (%)	2480 (55.1)	363 (58.1)	.17	752 (32.0)	34 (22.2)	.02
Hypertension, n (%)	2994 (66.5)	378 (60.5)	.003	1209 (51.4)	55 (35.9)	<.001
Chronic kidney disease, n (%)	575 (12.8)	85 (13.6)	.61	429 (18.3)	21 (13.7)	.19
Dialysis, n (%)	231 (5.1)	47 (7.5)	.02	159 (6.8)	7 (4.6)	.38
Current smoking, n (%)	396 (8.8)	58 (9.3)	.75	157 (6.7)	16 (10.5)	.11
Current alcohol, n (%)	660 (14.7)	89 (14.2)	.83	260 (11.1)	15 (9.8)	.73
Coronary artery disease, n (%)	1059 (23.5)	111 (17.8)	.002	430 (18.3)	11 (7.2)	.001
Previous disease						
Old myocardial infarction, n (%)	388 (8.6)	60 (9.6)	.46	215 (9.1)	12 (7.8)	.69
History of coronary intervention, n (%)	530 (11.8)	36 (5.8)	<.001	304 (12.9)	10 (6.5)	.03
History of coronary artery bypass graft, n (%)	120 (2.7)	17 (2.7)	>.99	66 (2.8)	4 (2.6)	>.99
Heart failure, n (%)	174 (3.9)	14 (2.2)	.06	66 (2.8)	5 (3.3)	.94
Stroke, n (%)	415 (9.2)	78 (12.5)	.01	253 (10.8)	17 (11.1)	>.99
Atrial fibrillation, n (%)	356 (7.9)	55 (8.8)	.49	169 (7.2)	8 (5.2)	.45
Arrhythmia, n (%)	453 (10.1)	63 (10.1)	>.99	229 (9.7)	12 (7.8)	.53
Valvular heart disease, n (%)	95 (2.1)	8 (1.3)	.22	117 (5.0)	8 (5.2)	>.99
Aortic disease, n (%)	136 (3.0)	14 (2.2)	.34	145 (6.2)	5 (3.3)	.20
Peripheral arterial disease, n (%)	146 (3.2)	11 (1.8)	.06	91 (3.9)	7 (4.6)	.83
Chronic pulmonary disease, n (%)	282 (6.3)	32 (5.1)	.30	206 (8.8)	9 (5.9)	.28
Active cancer, n (%)	1751 (38.9)	262 (41.9)	.16	798 (34.0)	34 (22.2)	.004
Charlson score, mean (SD)	3.2 (2.2)	3.8 (2.3)	<.001	2.1 (2.1)	1.6 (1.7)	<.001
Operative variables						
ESC[a]/ESA[b] surgical high risk, n (%)	1216 (27.0)	143 (22.9)	.03	524 (22.3)	38 (24.8)	.53
Emergency operation, n (%)	1167 (25.9)	318 (50.9)	<.001	483 (20.6)	83 (54.2)	<.001
General anesthesia, n (%)	3947 (87.7)	528 (84.5)	.03	2047 (87.1)	128 (83.7)	.27
Operation duration (hours), mean (SD)	3.7 (2.8)	3.1 (2.8)	<.001	3.0 (2.2)	2.7 (2.3)	.12
Packed red blood cell transfusion, n (%)	695 (15.4)	112 (17.9)	.13	0.5 (1.5)	1.1 (2.0)	<.001
Postoperative in-hospital events						
Type I myocardial infarction, n (%)	104 (2.3)	16 (2.6)	.81	8 (0.3)	2 (1.3)	.24
Coronary revascularization, n (%)	151 (3.4)	11 (1.8)	.04	28 (1.2)	2 (1.3)	>.99
Percutaneous coronary intervention, n (%)	134 (3.0)	8 (1.3)	.02	27 (1.1)	2 (1.3)	>.99
C-reactive protein level at discharge, mean (SD)	3.6 (4.0)	9.7 (8.8)	<.001	3.6 (4.4)	10.1 (8.7)	<.001
Medication at discharge, n (%)						
Beta blocker	1031 (22.9)	13 (2.1)	<.001	294 (12.5)	5 (3.3)	.001
Calcium channel blocker	1224 (27.2)	21 (3.4)	<.001	800 (34.0)	18 (11.8)	<.001

	Training data set			Test data set		
	No 30-day mortality (n=4501)	30-day mortality (n=625)	*P* value	No 30-day mortality (n=2350)	30-day mortality (n=153)	*P* value
Diltiazem	384 (8.5)	14 (2.2)	<.001	125 (5.3)	1 (0.7)	.02
Stain	1165 (25.9)	12 (1.9)	<.001	933 (39.7)	9 (5.9)	<.001
Metformin	497 (11.0)	26 (4.2)	<.001	474 (20.2)	7 (4.6)	<.001
Insulin	1127 (25.0)	335 (53.6)	<.001	636 (27.1)	73 (47.7)	<.001
Antiplatelet	1515 (33.7)	10 (1.6)	<.001	798 (34.0)	9 (5.9)	<.001
Renin angiotensin aldosterone system inhibitor	1105 (24.6)	20 (3.2)	<.001	677 (28.8)	9 (5.9)	<.001
Direct oral anticoagulant	211 (4.7)	3 (0.5)	<.001	6 (0.3)	0 (0.0)	>.99

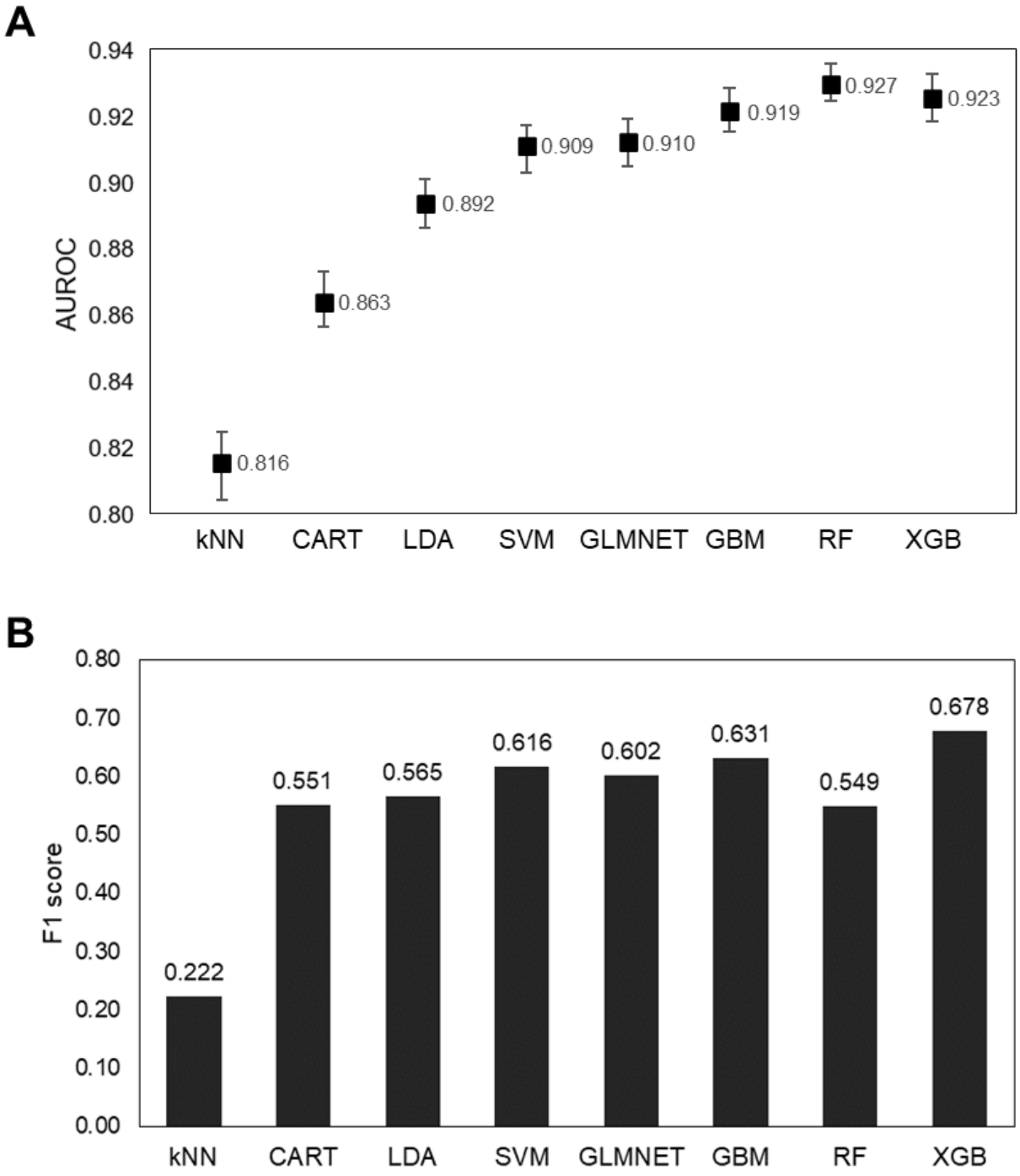
[a]ESC: European Society of Cardiology.

[b]ESA: European Society of Anaesthesiology

Development of a 30-Day Mortality Prediction Model

The probability of developing a 30-day mortality prediction model was explored using 8 machine learning algorithms. The hyperparameters optimized using grid search are summarized in Multimedia Appendix 7. The performance of each model is displayed using AUROC and AUPRC plots (Multimedia Appendix 8) along with various indexes (Multimedia Appendix 9). The performance of the kNN, CART, LDA, SVM, GLMNET, and GBM models was lower than that of the RF and XGB models. The RF and XGB models showed comparable performances. The AUROC of the RF model (0.927) was higher than that of the XGB model (0.923) in the training phase. However, the AUPRC of the RF model (0.747) was lower than that of the XGB model (0.763). Additionally, the F1 score and balanced accuracy of the XGB model (0.678 and 0.784) were higher than those of the RF model (0.549 and 0.695). When the models were comprehensively evaluated, the XGB model was selected as the best performing model for predicting the 30-day mortality of patients with MINS (Figure 2).

Figure 2. Performance comparison of each 30-day mortality prediction model with the range of (A) AUROC and (B) F1 score. AUROC: area under the receiver operating characteristic; CART: classification and regression trees; GBM: generalized boosted regression model; GLMNET: lasso/ridge/elastic net; kNN: k-nearest neighbors; LDA: linear discriminant analysis; RF: random forests; SVM: support vector machines; XGB: extreme gradient boosting.



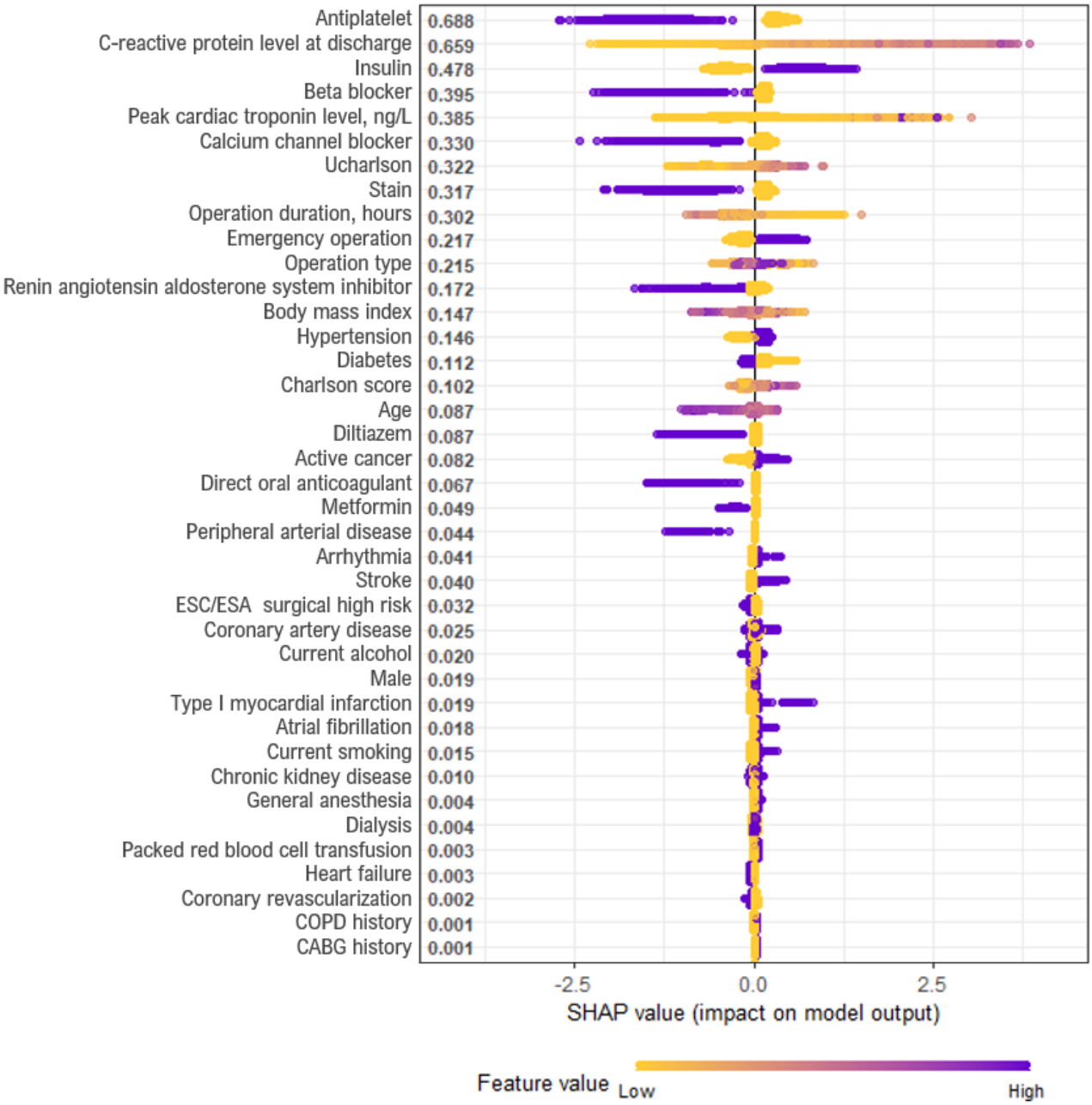
XGB 30-Day Mortality Prediction Model Interpretation

We tried to enable models to be actively accommodated by securing an interpretability and transparency. The importance of features in the XGB model is based on an algorithm that reduces based on the impurity index of the binary tree. The feature importance plot of the XGB 30-day mortality prediction model is shown in Multimedia Appendix 10. The top 5 features were C-reactive protein (CRP) level at discharge, antiplatelet prescription at discharge, peak cardiac troponin levels (ng/L), insulin prescription at discharge, and operation duration (hours).

The SHAP summary plot for the XGB models is shown in Figure 3. The XGB models determined that antiplatelet prescription at discharge was the most important variable, followed by CRP level at discharge, insulin prescription at discharge, beta blocker prescription at discharge, and peak cardiac troponin level (ng/L). According to the SHA*P* values of each feature, antiplatelet prescription at discharge was associated with a lower probability of death (left side of the

vertical dotted lines). Higher CRP levels at discharge were associated with a higher probability of death. Insulin prescription at discharge was associated with higher probability of death. Additionally, a SHAP dependence plot was used to explain how a single feature affects the output of the XGB prediction model (Multimedia Appendix 11).

Figure 3. SHAP summary plot of 30-day mortality prediction extreme gradient boosting model. According to the SHAP values of each feature, antiplatelet prescription at discharge (ie, purple dots) was associated with a lower probability of death (ie, the left side of the vertical dotted line). Higher C-reactive protein levels at discharge (ie, purple dots) were associated with a higher probability of death (ie, the right side of the vertical dotted line). Insulin prescription at discharge (ie, purple dots) was associated with a higher probability of death (ie, the right side of the vertical dotted line). CABG: coronary artery bypass graft; COPD: chronic obstructive pulmonary disease; ESA: European Society of Anaesthesiology; ESC: European Society of Cardiology; SHAP: Shapley Additive Explanations.



Lightening the Model Using Feature Selection

By reducing the number of variables required to use predictive models, we tried to make the model more acceptable in clinical practice. We used the recursive feature elimination (RFE) method to explore the relation between the number of features and performance. According to the RFE method, the accuracy of the model is best when the top 28 variables were used. However, the performance of the model was almost the same as when the top 10 variables were used (Multimedia Appendix 12). To minimize the number of variables input into the model, we observed the changes in performance while reducing the number of variables to 28, 10, and 5.

Light Model With 28 Variables

The list of the top 28 predictor variables chosen by the RFE method is shown in Multimedia Appendix 12. When the top 28 variables were used to train the model, the performance of the

XGB model had an accuracy of 0.926, AUPRC of 0.754, and F1 score of 0.652 (Multimedia Appendix 13). The AUROC was 0.925 (95% CI 0.919-0.931) in the training phase and 0.908 (95% CI 0.877-0.932) in the test phase. The AUROC of the model did not significantly decrease on the test data set (*P*=.22; Multimedia Appendix 14).

Light Model With 10 Variables

The top 10 variables used to train the XGB model were "crp_predc," "insulin_dc," "x_antiplt_dc," "peaktro," "ccb_dc," "emergencyop," "opduration," "statin_dc," "bb_dc," and "optype." The XGB model had an accuracy of 0.920, AUPRC of 0.708, and F1 score of 0.616 (Multimedia Appendix 13). The AUROC was 0.911 (95% CI 0.904-0.918) in the training phase and 0.904 (95% CI 0.874-0.93) in the test phase. The AUROC of the model did not significantly decrease on the test data set (*P*=.65; Multimedia Appendix 14).

Light Model With 10 Variables Chosen for Clinical Prediction

We made another model using 10 variables chosen for clinical prediction. Currently used treatments for MINS include dabigatran, a type of direct-acting oral coagulant [16], and potential treatments include antiplatelet agents and statins. We aimed to create a predictive model after excluding these drugs from the variables. The 10 chosen variables used were "crp_predc," "insulin_dc," "peaktro," "ccb_dc," "ccb_dc," "emergencyop," "opduration," "bb_dc," "optype," "x_raas_dc," and "metformin_dc." The XGB model had an accuracy of 0.916, AUPRC of 0.672, and F1 score of 0.587 (Multimedia Appendix 13). The AUROC was 0.894 (95% CI 0.887-0.902) in the training phase and 0.895 (95% CI 0.867-0.923) in the test phase. The AUROC of the XGB model did not significantly decrease on the test data set (*P*=.99; Multimedia Appendix 14).

Light Model With 5 Variables

Multimedia Appendix 12 shows that the prediction accuracy decreased by approximately 1.9% when the model used 5 variables compared to when the model used 28 variables. For users who have only a small amount of information about patients with MINS, we made a lighter model by selecting 5 variables based on the RFE's feature order. The top 5 variables used were "crp_predc," "insulin_dc," "x_antiplt_dc," "peaktro," and "ccb_dc." The XGB model had an accuracy of 0.907, AUPRC of 0.640, and F1 score of 0.505 (Multimedia Appendix 13). The AUROC was 0.890 (95% CI 0.882-0.898) in the training phase and 0.885 (95% CI 0.856-0.915) in the test phase. The AUROC of the model did not significantly decrease on the test data set (*P*=.80; Multimedia Appendix 14).

Development of a 1-Year Mortality Prediction Model

The AUROC of the 1-year mortality prediction XGB model was evaluated using the optimized hyperparameters eta=0.1, gamma=0, max tree depth=4, nround=100, colsample bytree=0.6, min child weight=1, and subsample=1. The AUROC of the model was 0.857 (95% CI 0.85-0.864) on the training data set and 0.794 (95% CI 0.756-0.826) on the test data set (Multimedia Appendix 15). The AUROC decreased on the test data set, and a statistically significant difference was observed compared to the AUROC of the training data set (*P*<.001). However, the prediction of the model is still valuable because the accuracy (0.95) on the test data set was above the no information rate (*P*=.001; Multimedia Appendix 16).

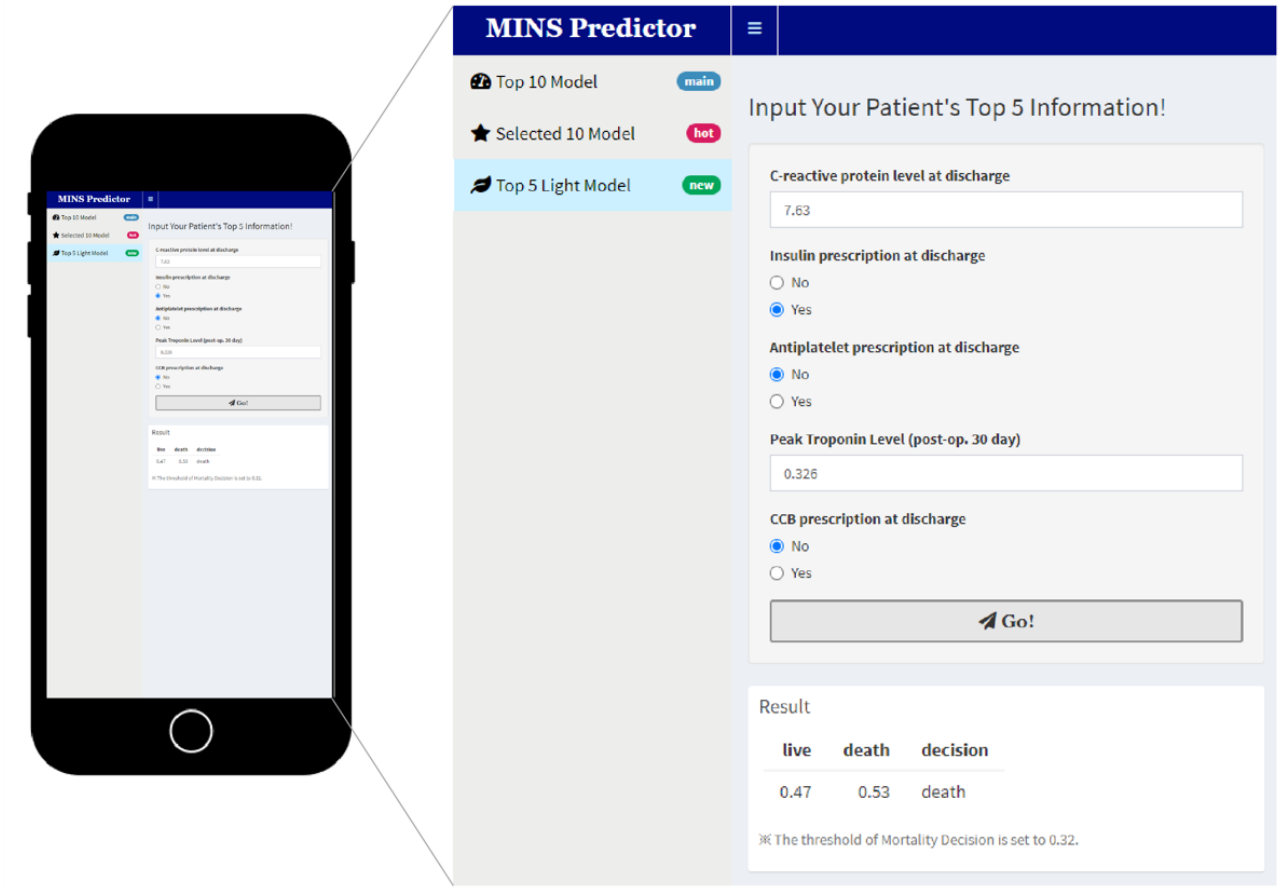
The feature importance plot of the 1-year mortality prediction model is shown in Multimedia Appendix 17. The top five features were the CRP level at discharge, peak cardiac troponin level (ng/L), operation duration (hours), antiplatelet prescription at discharge, and ucharlson score.

The SHAP summary plot for the models is shown in Multimedia Appendix 18. The XGB models determined that the CRP level at discharge was the most important variable, followed by the ucharlson score, antiplatelet prescription at discharge, insulin prescription at discharge, and operation duration (hours). According to the SHA*P* values of each feature, a higher CRP level at discharge and ucharlson score were associated with a higher probability of death. Antiplatelet prescription at discharge was associated with a lower probability of death, and insulin prescription at discharge was associated with a higher probability of death.

Development of an App With 30-Day Mortality Prediction XGB Model

The app, Leveraging R Shiny, was developed for practical use of the 30-day mortality prediction XGB model (Figure 4). Users can download the app for free via the public link [17]. Three versions of light models developed in this study are incorporated in the app: the top 10 features model, chosen 10 features model, and top 5 model. Each model was explored to find the optimal threshold for predicting patients at high risk of death. The optimized thresholds were applied to each model: 0.65 for the top 10 model; 0.53, chosen 10 model; and 0.68, top 5 model (Multimedia Appendix 19). Each user can choose a model type according to the type of variables that can be entered in a medical situation. A value for each variable corresponding to the target patient is entered and the Action button is pressed for probability output of the patient's demise in 30 days. After adjusting certain variable values, clinicians can observe changes in mortality and apply them to treatment decisions.

Figure 4. Internet app for predicting 30-day mortality of patients with MINS. CCB: calcium channel blockers; MINS: myocardial injury after noncardiac surgery.



Discussion

In this observational cohort study, we demonstrated the predictability of mortality in patients with MINS based on perioperative variables using a machine learning method.

Analysis of Model Performance Considering the Asymmetry of Data

To avoid overestimating the performance of the model, an imbalanced data set should be treated carefully when training a supervised classification machine learning model [18,19]. Along with accuracy, we wanted to interpret the performance of the model using indicators such as precision, recall, F1 score, AUPRC, and no information rate. In addition, for calibrating imbalanced data, four methods including oversampling, undersampling, both-sampling, and Random Over-Sampling Examples–sampling were carried out on the training data set, but the model's performance was significantly reduced when the model was applied in the test data set; therefore, these methods were not accepted (data not shown).

Comparison of 30-Day and 1-Year Preference Model Performance

We investigated why the 1-year prediction performance was lower than the 30-day prediction in this study. First, predicting the distant future is harder than predicting the near future. From a clinical perspective, although MINS has been reported to be associated with mortality up to 2 years after surgery, more clinical events that affect mortality are likely to take place as the duration of follow-up extends. Additionally, the observation period of the patients who made up the test data set (1.5 years) was shorter than that of the training data set (9.5 years). The observation period of the test data set may have been too short to reflect the characteristics of a patient who died within 1 year.

Consideration of the SHAP Values of the Charlson Scores

We observed different relations to SHA*P* values between the original Charlson Comorbidity Index (CCI) scores and the updated CCI scores. The original CCI score shows a moderate proportional relationship with the SHA*P* value. However, the updated CCI score shows that the SHA*P* value increased rapidly in the low scores and was then maintained (Multimedia Appendix 20). It is assumed that the updated CCI score has changed the weights of cardiomyopathy, peripheral vascular disease, and cerebrovascular disease from 1 to 0.

Clinical Implications

MINS is the most common medical complication directly related to mortality [13]. The rapid detection and appropriate management of MINS affects many patients at risk of mortality. The only treatment that was established in randomized trials was direct oral anticoagulants [16]. However, strengthening of cardiovascular drugs such as aspirin, statins, and few types of hypertension drugs have been reported to be linked to reduced mortality in patients with MINS [10,20]. Our results, show that the prescription of cardiovascular drugs such as antiplatelet

agents, antihypertensive drugs, and statins at discharge are effective in predicting MINS mortality. The CRP level as a degree of inflammation is linked with the prognosis of coronary artery disease [21] and shows a strong impact on the model, which is consistent with previous studies. Therefore, our findings regarding the mortality of patients with MINS may be predicted based on perioperative variables, suggesting the possibility of reducing the mortality of patients with MINS by correction of perioperative variables.

We were able to reduce the number of variables to 5 with affordable loss in performance. Using only 5 variables, it is possible to predict the mortality of patients with MINS with 90.7% accuracy. A smaller number of variables in the prediction model indicates that it is highly likely to be used in other hospitals. Hence, we see this result as an important clinical implication.

Limitations of the Study

Our study has a few limitations. First, model validation was performed using a test data set having a different time window from that of the data set used for training and internal validation. As a study using observational data collected in a single institution, our predictive models may have limited generalizability. Using a data set of patients with MINS visiting different institutions over the same period would allow for more appropriate external validation.

Second, our results might have been affected by selection bias and confounding factors. Postoperative hs-cTn measurements were not routine and optionally performed in patients with specific cardiovascular risks. Consequently, the possibility of selection bias may exist and should be considered if the user wants to apply the model in clinical practice.

Third, after confirming that mortality can be predicted using observational data, we created and released a mobile app for users. However, the predictive model developed in this study cannot be immediately used in routine clinical practice. We plan to conduct further research to measure the applicability of the model in clinical practice.

Conclusions

We have confirmed that a 30-day mortality prediction model can be developed for patients with MINS using observational clinical data. The XGB algorithm outperformed the LDA, kNN, CART, SVM, GLMNET, RF, and GBM machine learning algorithms. To maximize the applicability of the prediction model in clinic settings, we observed that the number of variables that need to be input into the model can be reduced to 5 while preserving the performance of the model. For more robust evidence, a randomized clinical trial is required to address the variables explored in this study. However, this study is the first to report mortality predictability in patients with MINS using machine learning.

Acknowledgments

SJS and JP contributed equally to this study as co-first authors. KY and RWP contributed equally to this study as corresponding authors. This research was funded by the Bio Industrial Strategic Technology Development Program (20003883, 20005021) funded by the Ministry of Trade, Industry & Energy (Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (grant HR16C0001). This research was also supported by Healthcare AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea funded by the Ministry of Science and ICT (No. 1711120339).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The clinical data warehouse of Samsung Medical Center, named DARWIN-C.
[DOCX File , 347 KB - medinform_v9i10e32771_app1.docx]

Multimedia Appendix 2
Name and meaning of features in original and test data sets.
[DOCX File , 19 KB - medinform_v9i10e32771_app2.docx]

Multimedia Appendix 3
Performance indicators for evaluating machine learning models.
[DOCX File , 14 KB - medinform_v9i10e32771_app3.docx]

Multimedia Appendix 4
Age distribution and sex of the patients in the two data sets.
[DOCX File , 185 KB - medinform_v9i10e32771_app4.docx]

Multimedia Appendix 5

Descriptive analysis of surgery type.
[DOCX File , 189 KB - medinform_v9i10e32771_app5.docx]

Multimedia Appendix 6
Surgery type and mortality.
[DOCX File , 17 KB - medinform_v9i10e32771_app6.docx]

Multimedia Appendix 7
The optimal parameters of the machine learning algorithm.
[DOCX File , 17 KB - medinform_v9i10e32771_app7.docx]

Multimedia Appendix 8
Area under the receiver operating characteristic and area under the precision and recall curve plots of each model in predicting 30-day mortality.
[DOCX File , 409 KB - medinform_v9i10e32771_app8.docx]

Multimedia Appendix 9
Performance indexes of machine learning models predicting 30-day mortality of patients with myocardial injury after noncardiac surgery.
[DOCX File , 26 KB - medinform_v9i10e32771_app9.docx]

Multimedia Appendix 10
Importance of features in the extreme gradient boosting 30-day mortality prediction model.
[DOCX File , 154 KB - medinform_v9i10e32771_app10.docx]

Multimedia Appendix 11
Shapley Additive Explanations dependence plots for top 10 features of 30-day mortality prediction model.
[DOCX File , 115 KB - medinform_v9i10e32771_app11.docx]

Multimedia Appendix 12
Recursive feature elimination graph based on accuracy.
[DOCX File , 31 KB - medinform_v9i10e32771_app12.docx]

Multimedia Appendix 13
Performance indexes of extreme gradient boosting models predicting 30-day mortality of patients with myocardial injury after noncardiac surgery with top 28, top 10, chosen 10, and top 5 variables.
[DOCX File , 20 KB - medinform_v9i10e32771_app13.docx]

Multimedia Appendix 14
Area under the receiver operating characteristic and area under the precision and recall curve plots of each model predicting 30-day mortality with top 28, top 10, chosen 10, and top 5 variables.
[DOCX File , 210 KB - medinform_v9i10e32771_app14.docx]

Multimedia Appendix 15
Area under the receiver operating characteristic and area under the precision and recall curve plots of the extreme gradient boosting model predicting 1-year mortality.
[DOCX File , 68 KB - medinform_v9i10e32771_app15.docx]

Multimedia Appendix 16
Performance indexes of extreme gradient boosting model predicting 1-year mortality of patients with myocardial injury after noncardiac surgery.
[DOCX File , 15 KB - medinform_v9i10e32771_app16.docx]

Multimedia Appendix 17
Importance of features in the extreme gradient boosting 1-year mortality prediction model.
[DOCX File , 158 KB - medinform_v9i10e32771_app17.docx]

Multimedia Appendix 18
Shapley Additive Explanations summary plot of 1-year mortality prediction extreme gradient boosting model.
[DOCX File , 158 KB - medinform_v9i10e32771_app18.docx]

Multimedia Appendix 19
Optimized threshold and final performance of models.
[DOCX File , 13 KB - medinform_v9i10e32771_app19.docx]

Multimedia Appendix 20
Comparison of the Shapley Additive Explanations value with the Charlson Comorbidity Index score and updated Charlson Comorbidity Index score.
[DOCX File , 327 KB - medinform_v9i10e32771_app20.docx]

References

1. Puelacher C, Lurati Buse G, Seeberger D, Sazgary L, Marbot S, Lampart A, BASEL-PMI Investigators. Perioperative myocardial injury after noncardiac surgery: incidence, mortality, and characterization. Circulation 2018 Mar 20;137(12):1221-1232. [doi: 10.1161/CIRCULATIONAHA.117.030114] [Medline: 29203498]
2. Smilowitz NR, Redel-Traub G, Hausvater A, Armanious A, Nicholson J, Puelacher C, et al. Myocardial injury after noncardiac surgery: a systematic review and meta-analysis. Cardiol Rev 2019;27(6):267-273 [FREE Full text] [doi: 10.1097/CRD.0000000000000254] [Medline: 30985328]
3. Sessler DI, Devereaux PJ. Perioperative troponin screening. Anesth Analg 2016 Aug;123(2):359-360. [doi: 10.1213/ANE.0000000000001450] [Medline: 27331782]
4. van Waes JAR, van Klei WA, Wijeysundera DN, van Wolfswinkel L, Lindsay TF, Beattie WS. Association between intraoperative hypotension and myocardial injury after vascular surgery. Anesthesiology 2016 Jan;124(1):35-44 [FREE Full text] [doi: 10.1097/ALN.0000000000000922] [Medline: 26540148]
5. Schacham YN, Cohen B, Bajracharya GR, Walters M, Zimmerman N, Mao G, et al. Mild perioperative hypothermia and myocardial injury: a retrospective cohort analysis. Anesth Analg 2018 Dec;127(6):1335-1341. [doi: 10.1213/ANE.0000000000003840] [Medline: 30300173]
6. van Lier F, Wesdorp FHIM, Liem VGB, Potters JW, Grüne F, Boersma H, et al. Association between postoperative mean arterial blood pressure and myocardial injury after noncardiac surgery. Br J Anaesth 2018 Jan;120(1):77-83 [FREE Full text] [doi: 10.1016/j.bja.2017.11.002] [Medline: 29397140]
7. Turan A, Cohen B, Rivas E, Liu L, Pu X, Maheshwari K, et al. Association between postoperative haemoglobin and myocardial injury after noncardiac surgery: a retrospective cohort analysis. Br J Anaesth 2021 Jan;126(1):94-101. [doi: 10.1016/j.bja.2020.08.056] [Medline: 33039122]
8. Kwon J, Park J, Lee S, Lee JH, Min JJ, Kim J, et al. Pre-operative anaemia and myocardial injury after noncardiac surgery: a retrospective study. Eur J Anaesthesiol 2021 Jun 01;38(6):582-590. [doi: 10.1097/EJA.0000000000001421] [Medline: 33399380]
9. Lee S, Yang K, Park J, Lee JH, Min JJ, Kwon J, et al. Association between high body mass index and mortality following myocardial injury after noncardiac surgery. Anesth Analg 2021 Apr 01;132(4):960-968. [doi: 10.1213/ANE.0000000000005303] [Medline: 33323785]
10. Park J, Kim J, Lee S, Lee JH, Min JJ, Kwon J, et al. Postoperative statin treatment may be associated with improved mortality in patients with myocardial injury after noncardiac surgery. Sci Rep 2020 Jul 15;10(1):11616. [doi: 10.1038/s41598-020-68511-3] [Medline: 32669686]
11. Oh AR, Park J, Lee S, Kim J, Lee JH, Min JJ, et al. Elevated high-sensitivity C-reactive protein concentrations may be associated with increased postdischarge mortality in patients with myocardial injury after noncardiac surgery: A retrospective observational study. Eur J Anaesthesiol 2021 Mar 01;38(Suppl 1):S33-S40. [doi: 10.1097/EJA.0000000000001409] [Medline: 33399373]
12. Devereaux PJ, Szczeklik W. Myocardial injury after non-cardiac surgery: diagnosis and management. Eur Heart J 2020 May 01;41(32):3083-3091. [doi: 10.1093/eurheartj/ehz301] [Medline: 31095334]
13. Kristensen SD, Knuuti J. New ESC/ESA Guidelines on non-cardiac surgery: cardiovascular assessment and management. Eur Heart J 2014 Sep 14;35(35):2344-2345. [doi: 10.1093/eurheartj/ehu285] [Medline: 25104785]
14. Mahajan VS, Jarolim P. How to interpret elevated cardiac troponin levels. Circulation 2011 Nov 22;124(21):2350-2354. [doi: 10.1161/CIRCULATIONAHA.111.023697] [Medline: 22105197]
15. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct;2(10):749-760 [FREE Full text] [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]

16. Devereaux PJ, Duceppe E, Guyatt G, Tandon V, Rodseth R, Biccard BM, MANAGE Investigators. Dabigatran in patients with myocardial injury after non-cardiac surgery (MANAGE): an international, randomised, placebo-controlled trial. Lancet 2018 Jun 09;391(10137):2325-2334. [doi: 10.1016/S0140-6736(18)30832-8] [Medline: 29900874]
17. Shin SJ, Yang KM. MINS Predictor. URL: https://bit.ly/3zepij6 [accessed 2021-08-09]
18. Karajizadeh M, Nasiri M, Yadollahi M, Zolfaghari AH, Pakdam A. Mortality prediction from hospital-acquired infections in trauma patients using an unbalanced dataset. Healthc Inform Res 2020 Oct;26(4):284-294 [FREE Full text] [doi: 10.4258/hir.2020.26.4.284] [Medline: 33190462]
19. Symum H, Zayas-Castro JL. Prediction of chronic disease-related inpatient prolonged length of stay using machine learning algorithms. Healthc Inform Res 2020 Jan;26(1):20-33 [FREE Full text] [doi: 10.4258/hir.2020.26.1.20] [Medline: 32082697]
20. Foucrier A, Rodseth R, Aissaoui M, Ibanes C, Goarin J, Landais P, et al. The long-term impact of early cardiovascular therapy intensification for postoperative troponin elevation after major vascular surgery. Anesth Analg 2014 Nov;119(5):1053-1063. [doi: 10.1213/ANE.0000000000000302] [Medline: 24937347]
21. Ridker PM, MacFadyen JG, Everett BM, Libby P, Thuren T, Glynn RJ, CANTOS Trial Group. Relationship of C-reactive protein reduction to cardiovascular event reduction following treatment with canakinumab: a secondary analysis from the CANTOS randomised controlled trial. Lancet 2018 Jan 27;391(10118):319-328. [doi: 10.1016/S0140-6736(17)32814-3] [Medline: 29146124]

Abbreviations

AUPRC: area under the precision and recall curve
AUROC: area under the receiver operating characteristic
CART: classification and regression trees
CCI: Charlson Comorbidity Index
CRP: C-reactive protein
GBM: generalized boosted regression model
GLMNET: lasso/ridge/elastic net
kNN: k-nearest neighbors
LDA: linear discriminant analysis
MINS: myocardial injury after noncardiac surgery
RF: random forests
RFE: recursive feature elimination
SHAP: Shapley Additive Explanations
SMC-TINCO: Samsung Medical Center Troponin in Noncardiac Operation
SVM: support vector machines
XGB: extreme gradient boosting

Edited by G Eysenbach; submitted 09.08.21; peer-reviewed by Y Kim; comments to author 30.08.21; revised version received 31.08.21; accepted 20.09.21; published 14.10.21.

Please cite as:
Shin SJ, Park J, Lee SH, Yang K, Park RW
Predictability of Mortality in Patients With Myocardial Injury After Noncardiac Surgery Based on Perioperative Factors via Machine Learning: Retrospective Study
JMIR Med Inform 2021;9(10):e32771
URL: https://medinform.jmir.org/2021/10/e32771
doi:10.2196/32771
PMID:34647900

Original Paper

Ensemble Learning-Based Pulse Signal Recognition: Classification Model Development Study

Jianjun Yan[1], D; Xianglei Cai[1], MSc; Songye Chen[1], MSc; Rui Guo[2], MD; Haixia Yan[2], MD; Yiqin Wang[2], MD

[1]Institute of Intelligent Perception and Diagnosis, School of Mechanical and Power Engineering, East China University of Science and Technology, Shanghai, China

[2]Shanghai Key Laboratory of Health Identification and Assessment, Laboratory of Traditional Chinese Medicine for Diagnostic Information, Shanghai University of Traditional Chinese Medicine, Shanghai, China

Corresponding Author:
Jianjun Yan, D
Institute of Intelligent Perception and Diagnosis
School of Mechanical and Power Engineering
East China University of Science and Technology
130 Meilong Road
Shanghai, 200237
China
Phone: 86 21 64252074
Email: jjyan@ecust.edu.cn

Abstract

Background: In pulse signal analysis and identification, time domain and time frequency domain analysis methods can obtain interpretable structured data and build classification models using traditional machine learning methods. Unstructured data, such as pulse signals, contain rich information about the state of the cardiovascular system, and local features of unstructured data can be extracted and classified using deep learning.

Objective: The objective of this paper was to comprehensively use machine learning and deep learning classification methods to fully exploit the information about pulse signals.

Methods: Structured data were obtained by using time domain and time frequency domain analysis methods. A classification model was built using a support vector machine (SVM), a deep convolutional neural network (DCNN) kernel was used to extract local features of the unstructured data, and the stacking method was used to fuse the above classification results for decision making.

Results: The highest average accuracy of 0.7914 was obtained using only a single classifier, while the average accuracy obtained using the ensemble learning approach was 0.8330.

Conclusions: Ensemble learning can effectively use information from structured and unstructured data to improve classification accuracy through decision-level fusion. This study provides a new idea and method for pulse signal classification, which is of practical value for pulse diagnosis objectification.

(JMIR Med Inform 2021;9(10):e28039) doi:10.2196/28039

KEYWORDS

wrist pulse; ensemble learning; support vector machine; deep convolutional neural network; pulse signal; machine learning; traditional Chinese medicine; pulse classification; pulse analysis; fully connected neural network; synthetic minority oversampling technique; feature extraction

Introduction

A pulse signal contains a large amount of pathological and physiological information [1,2], and the signal characteristics are closely related to diseases (hypertension, atherosclerosis, etc), especially cardiovascular disease (CVD) and physiological parameters (pulse wave velocity, blood pressure, etc) [3,4]. Therefore, pulse analysis is widely used for cardiovascular function assessment and noninvasive early diagnosis of CVD and related complications [5]. It is a convenient, noninvasive, and effective diagnostic method that is widely used in traditional Chinese medicine (TCM). In recent years, smart wearable devices have become increasingly popular, allowing individuals

to monitor their own pulse status. However, the important information contained in the pulse signal requires a highly experienced TCM practitioner to make a diagnosis, which is highly variable [6]. In Chinese medicine, pulses are classified into 28 single-pulse types based on 4 major elements: pulse depth, pulse rate, pulse shape, and pulse intensity [6,7]. A patient's pulse may be a combination of several single-pulse types, that is, a compound pulse [8], and a compound pulse may carry more physiological information and be more difficult to distinguish and identify. For example, a slippery pulse and a flat pulse are the main pulse types in healthy people; a thready pulse may be due to overexertion and deficiency of qi and blood; a stringy pulse may be related to liver disorders; a thready, slippery pulse may be related to colds; a thready, stringy pulse may be related to kidney disorders; and a stringy, slippery pulse may be related to coughing, dizziness, and weakness.

Practitioners of Chinese medicine make a diagnosis by touching the patient's wrist and feeling the patient's pulse with their fingers for several minutes to determine the patient's pulse type through experience and make medical decisions accordingly.

Using deep learning or machine learning methods, pulse types can be better classified to help medical practitioners with diagnosis. For individuals, without medical background and experience, the pulse types obtained can also be collected and analyzed by wearable devices to obtain a preliminary understanding of their physical condition and can better prevent CVD. There is already a good deal of scholarly research related to the classification of pulse types. Xu et al [9] used Lempel-Ziv complexity analysis to detect arrhythmic pulses. This approach was applied on 140 clinic pulses for detecting 7 pulse patterns. Zhang et al [10] referred to the edit distance with real penalty (ERP) and the progress in k-nearest-neighbor (KNN) classifiers using an ERP-based KNN classifier on the classification of pulse waveforms. Garmaev et al [11] used cluster analysis of the time parameters of a pulse signal to classify pulses. After clustering, the data were evaluated using the nonparametric Kruskel-Wallis test. Li et al [12] used five CVD and complications extracted from medical records as classification criteria. This convolutional neural network (CNN) could extract stronger features for pulse signals. Huang et al [13] developed a high-dimensional pulse classification method to improve pulse diagnosis accuracy. They extracted 71 pulse features from the time, spatial, and frequency domains to cover as much pulse information as possible.

However, most of the above methods extract structured data of pulse signals and are suitable for traditional machine learning models; however, unstructured data, such as pulse signals, contain rich information. Taking advantage of different pulse signal analysis methods and combining machine learning and deep learning methods to build pulse signals for classification models can help TCM practitioners make better pulse diagnoses and help smart wearable devices more accurately assess the human health status.

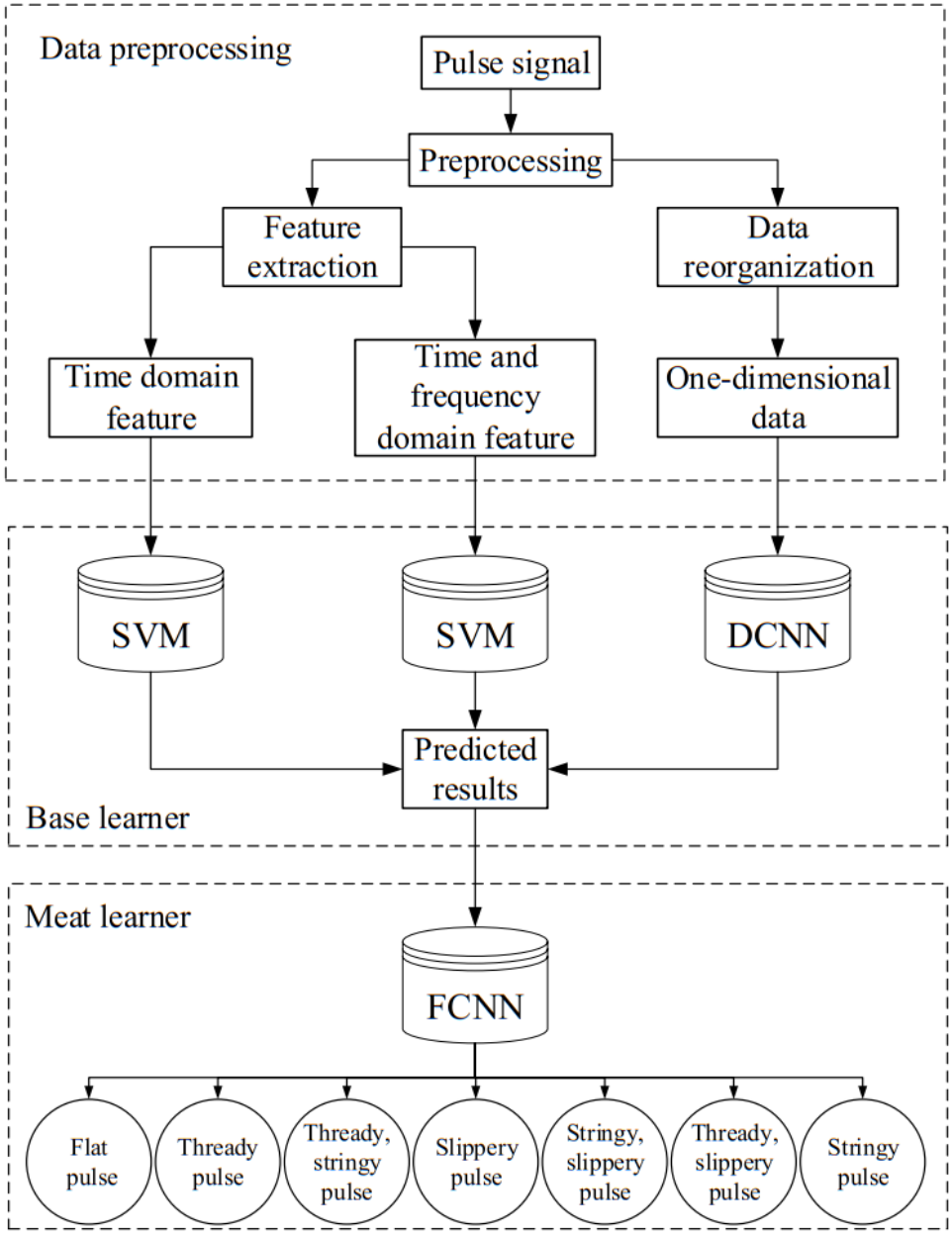
The objective of this paper was to comprehensively use machine learning and deep learning classification methods to fully exploit the information about pulse signals.

Methods

System Flow

A deep convolutional neural network (DCNN)- and SVM-based stacking network (DSSN) selected suitable algorithms to build classification models based on the structured and unstructured data extracted by different analysis methods and integrated them using the stacking method; the overall architecture is shown in Figure 1. First, the pulse signal was preprocessed to extract the feature parameters in time and frequency domains, and the data were reorganized for the pulse signal to prepare the data for the training of the base learners. An SVM and a DCNN were selected as the base learners to build the classification models corresponding to the three analysis methods, and the output results were combined together to form a new data set. Finally, the newly generated data set was used to train the meta-learner so as to build a DSSN pulse signal integrated classification model.

Figure 1. DSSN flowchart. During data preprocessing, time and frequency domain features of the pulse signal were extracted, and all pulse data were organized to the same length. Time and frequency domain features were separately trained by an SVM to obtain prediction results. One-dimensional data using the DCNN were used to obtain prediction results. Finally, the pulse-type prediction results of the three methods were integrated by an FCNN. SVM: support vector machine; DCNN: deep convolutional neural network; DSSN: DCNN- and SVM-based stacking network; FCNN: fully connected neural network.



Data

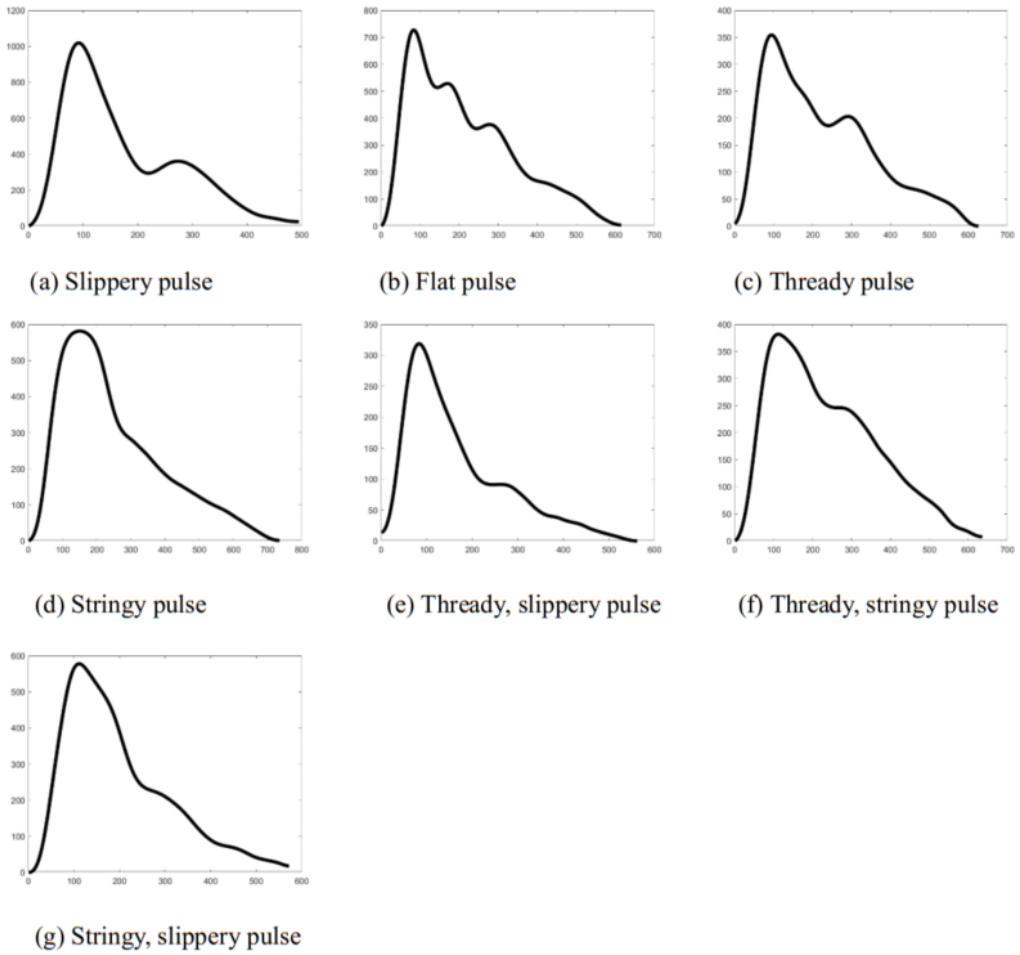
The experimental data in this paper were provided by the Four Diagnostic Information Comprehensive Research Laboratory of Shanghai University of Traditional Chinese Medicine, which included 7 types of pulse data (4 single pulses and 3 compound pulses), with a total sample size of 1812 cases; the specific pulse types and numbers are shown in Table 1. The acquisition device was a Z-BOX I pulse acquisition instrument with a sampling frequency of 720 Hz to acquire the pulse signals at the optimal pulse-taking pressure with an acquisition time of 60 s. Two or more TCM experts classified the collected pulse signals using their experience, and the pulse type was determined only when the majority of the experts agreed on the classification.

Table 1. Types of pulse signals and data size. There were 4 single pulses and 3 compound pulses. A total of 1812 cases were collected. After balancing the sample data, the total number of samples reached 4355.

Pulse type code	Name of pulse type	Sample size (n)	Balanced sample size (n)
1	Slippery pulse	221	637
2	Flat pulse	96	649
3	Thready pulse	92	607
4	Stringy pulse	657	630
5	Thready, slippery pulse	202	583
6	Thready, stringy pulse	325	614
7	Stringy, slippery pulse	219	635

The experimental data were first preprocessed, the samples were filtered and noise reduced, single-cycle segmentation of the pulse signal was performed, and the average single cycle was taken to represent the pulse signal; seven types of pulse data are shown in Figure 2. These data set also suffered from sample imbalance, and the synthetic minority oversampling technique (SMOTE) algorithm was used to equalize the data set. For data sets with sample imbalance, the basic approaches are oversampling and downsampling (ie, copying a small number of samples and removing a larger number of samples), but both pose problems: copying samples can easily make the model overfit, while removing samples can lead to a smaller number of samples. To solve such problems, the sampling SMOTE algorithm, which synthesizes new data, can compensate for the sample imbalance, while trying to avoid overfitting [14].

Figure 2. Seven types of pulse. Single pulses had four types: slippery, flat, thready, and stringy. Compound pulses had three types: thready slippery, thready stringy, and stringy slippery.



The number of samples in the equalized data set was 4355 in total after removing samples with failed acquisition and obvious errors in the waveform in the data set. The number of samples in each category is shown in Table 1. The SMOTE parameters were set as $k = 5$; sample multiplicity $N = 7$ for the flat, thready pulse; sample multiplicity N = 3 for the slippery, thready and slippery, and stringy and slippery pulses; and sample multiplicity $N = 2$ for the thready, stringy pulse.

Time Domain Feature Extraction

The time domain analysis method focuses on the waveform of the pulse signal in a typical cycle [15], defining the characteristic points with physiological and pathological significance and then extracting the corresponding characteristic parameters. In a single-cycle pulse signal, the peak and trough points of the waveform have certain physiological significance, and there are seven main feature points, including the start and end points, as shown in the marked A-G in Figure 3. The reference points and significance of the pulse waveform are shown in Table 2. The time domain feature extraction could be automatically performed using the algorithm of signal processing.

Figure 3. Basic information about a single-cycle pulse signal. The amplitude information about the pulse signal is represented vertically, and the time information is represented horizontally.

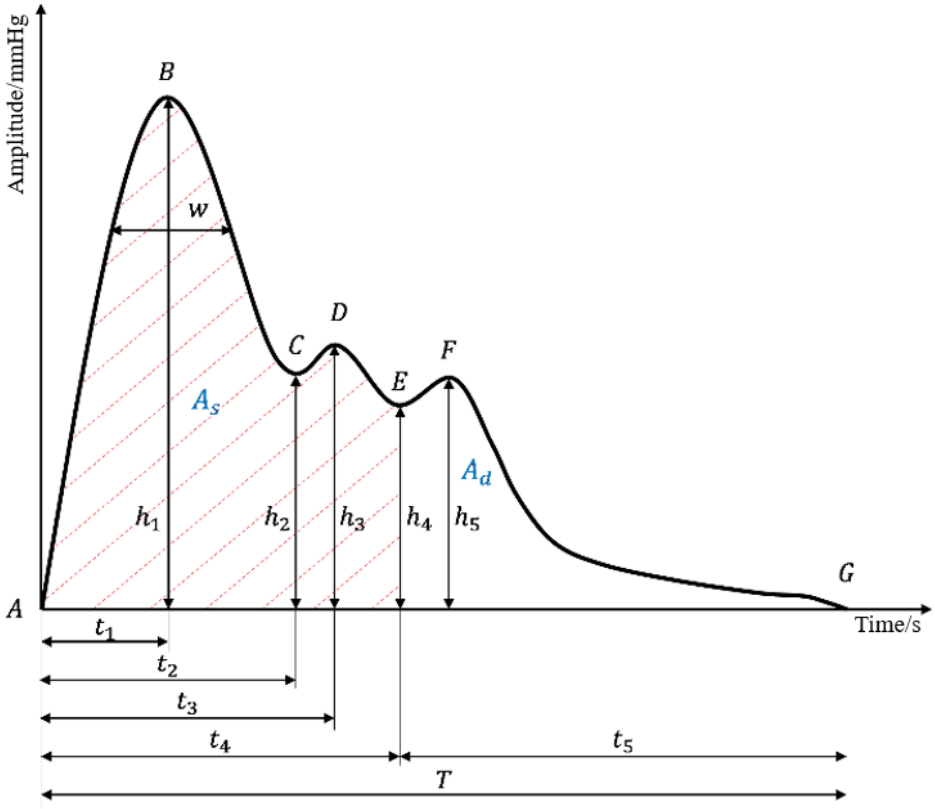


Table 2. Physiological significance of pulse signal reference points. These points include the start and end points of the pulse signal, as well as the extreme points of the pulse signal, all of which reflect to some extent the physiological information about the human body.

Reference point	Meaning
Systolic waveform	
A	Start point
B	Main wave crest
C	Main wave gap
D	Rebattling the former wave crest
Diastolic waveform	
E	Descending the middle gorge
F	Rebattling the wave
G	End point

In this paper, a total of 23 time domain features of the pulse signal were extracted, including 1 slope feature, 2 area features, 6 amplitude features, 6 time features, and 8 proportional features; the feature parameters and their specific meanings are shown in Table 3. When extracting the pulse waveform parameters, the ratio between different amplitudes was added as a feature in order to better reflect the waveform characteristics, such as h_2/h_1 and h_4/h_1, because of the large differences between different pulse waveforms. Similarly, the ratio between time parameters was increased to better distinguish between different types of pulse signals.

Table 3. Time domain features of the pulse signal. These features include 1 slope feature, 2 area features, 6 magnitude features, 6 time features, and 8 proportional features.

Feature type	Feature parameter	Feature name
Slope	k	Main wave slope
Area	A_s	Systolic area
	A_d	Diastolic area
Magnitude	h_1	Main wave amplitude
	h_2	Main wave gorge amplitude
	h_3	Wave front dicrotic amplitude
	h_4	Dicrotic notch amplitude
	h_5	Dicrotic wave amplitude
	w	1/3 pulse width
Time	t_1	Main wave phase
	t_2	Main wave gorge phase
	t_3	Wave front dicrotic phase
	t_4	Dicrotic notch phase
	t_5	Dicrotic wave phase
	T	Pulse cycle
Proportion	t_1/T	Time ratio
	t_1/t_4	Time ratio
	t_5/t_4	Time ratio
	w/T	Pulse width cycle ratio
	h_2/h_1	Main wave gorge main Wave amplitude ratio
	h_4/h_1	Dicrotic notch main wave amplitude ratio
	h_5/h_1	Dicrotic wave main wave amplitude ratio
	A_s/A_d	Systolic:diastolic area ratio

Time and Frequency Domain Feature Extraction

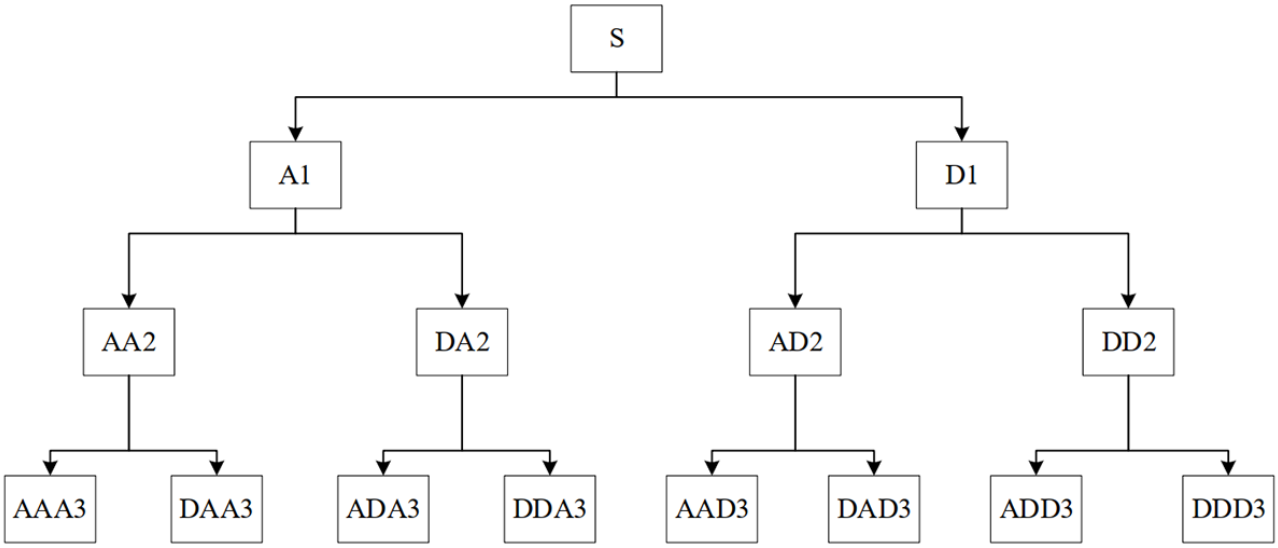
Wavelet packet analysis is an effective signal analysis method that uses different wavelet bases for signal decomposition and has a greater advantage in analyzing nonstationary signals. As a time frequency analysis method, wavelet packet analysis can zoom in on both time domain information and frequency domain information and has excellent time frequency local analysis capability. At present, it is applied to the analysis and identification of pulse signals, with good results [16]. As shown in Figure 4 as a schematic diagram of wavelet packet decomposition, wavelet packet analysis further decomposed the high-frequency band, while decomposing the low-frequency band, which improved the time-frequency resolution of the pulse signal. The low-frequency profile and high-frequency details of the wavelet packet decomposition, $u_n(t)$, at different frequencies is defined as follows:

where t is time, k is the time translation factor, and g_k and h_k have an orthogonal relationship, that is, $g_k = (-1)^k h_{1-k}$. Defining the number of coefficients in the j layer as n_j, the energy of the j-th layer is as follows:

The sampling frequency of the pulse signal was 720 Hz, and the energy of the pulse was mainly concentrated in the frequency band within 10 Hz. Therefore, the energy characteristics of the pulse signal in different frequency bands were extracted by 8-layer wavelet packet decomposition. For wavelet feature extraction of biological signals, sym8 outperforms Haar, dB2, and dB4 in terms of the performance index, specificity, sensitivity, accuracy, time delay, and quality assessment of wavelets [17]. A sym8 wavelet has better regularity and symmetry, which can reduce the phase distortion caused by the calculation. Combined with the above analysis, this paper used a sym8 wavelet to decompose the pulse signal by 8-layer wavelet packets to obtain 256-dimensional energy features. The

time and frequency domain features could be automatically extracted by the algorithm of signal processing.

Figure 4. Schematic diagram of wavelet packet decomposition.



Classification Methods

The pulse signal features extracted by using time domain and time frequency domain analyses are structured data, and unstructured data are obtained by using data reorganization to obtain one-dimensional data of the pulse signal. For structured data, a machine learning algorithm was applied; for unstructured data, a deep learning algorithm was used to train the classification model. Finally, all results were integrated using the stacking method. The algorithms used were as follows:

1. SVM: As a base learner, this method was used to train time domain and time frequency domain data. The penalty parameter C was set as 2.0, the kernel was set as rbf, and gamma was set as 3.0. The inputs to the SVM classifier were time domain features and time frequency domain features of the pulse. The outputs were the classification results of seven pulse types.

2. DCNN: As a base learner, the pulse signal whose length was 800 was the input to the DCNN classifier, and the outputs were the classification results of seven pulse types. The network used in this paper had three convolutional layers. The first layer was set as filters = 5 and kernel size = 11. The second layer was set as filters = 25 and kernel size = 9. The third layer was set as filters = 100 and kernel size = 10. The dense layer was set as units = 7 and activation = softmax.

3. FCNN: As a meta-learner, the input was results of the three base learners and the labels were the same as raw data. The outputs were classification results of the seven pulse types after stacking. The network had four dense layers. The first and second layers were set as units = 1024 and activation = relu. The third layer was set as units = 512 and activation = relu. The last layer was set as units = 7 and activation = softmax.

In this paper, deep neural networks, including the FCNN, were implemented by TensorFlow (Google) as the back-end Keras framework, the loss function was chosen as cross-entropy, the batch size of the DCNN was set to 8, the batch size of the FCNN was set to 32, the number of iterations was 1000, the initial learning rate was 0.001, the stochastic gradient descent (SGD) optimization algorithm was used, the momentum was 0.9, the weight recession was 0.0001, the dropout parameter was 0.5, and the ratio between the training set, validation set, and test set was 6:2:2. To avoid the overfitting phenomenon caused by network training, the early stop strategy was used. When the loss of the neural network on the validation set did not decrease within 10 cycles, the training was stopped and the model with the smallest loss on the validation set was selected as the final training result. The CPU used for the experiments was an Intel Core i7-8700K with 32 GB of memory and an NVIDIA Tesla V100 GPU graphics card.

The data set in this paper contained multiple pulse categories, so it was necessary to combine the classification results of each category for judgment. In this paper, macroaverage was used as the judging index, and the average accuracy and average recall were calculated as follows:

Results

SVM Experimental Results

In training the SVM model, the radial basis function (RBF) was chosen as the kernel function, in which the main parameters included the penalty coefficient C and gamma. c indicated the degree of acceptance of error; the larger the value of C, the less the classification error was allowed to occur during training, and the selection of appropriate C could suppress the overfitting phenomenon of the model. Here, gamma was a parameter of the RBF, which was used to adjust the range of action of the model support vector.

To find the best parameter that made the best classification of the model, this paper used the grid search method to determine optimal parameter values. When training the classification model with time domain features and time frequency domain features, the parameter range of C was set to 1-50, with a step size of 1; the parameter range of gamma was 1-50, with a step size of 0.5;

and default values were used for the rest of the parameters. The results of the optimal classification model are shown in Table 4.

Table 4. Classification results of time and time frequency domain features with an SVM[a]. The average accuracy rate is the percentage of all pulse type classifications that are correct. The average recall rate is the ratio of the correct pulse type in the classification result to the pulse type in the sample. Accuracy is the average of the accuracy of each of the 7 pulse types.

Classification model	Average accuracy rate (%)	Average recall rate (%)	Accuracy (%)
Time domain feature+SVM	79.2	76.2	76.1
Time and frequency domain feature+SVM	74.6	72.9	72.8

[a]SVM: support vector machine.

As can be seen from Table 4, the time domain classification model had a higher accuracy than the time frequency domain classification model with the same classifier, reaching 76.1%, which was 3.3% higher than the time frequency domain classification model. Meanwhile, the flat accuracy rate and the average recall rate of the time domain classification model were 79.2% and 76.2%, respectively, which were higher than those of the time frequency domain classification model.

DCNN Experimental Results

To verify the classification performance of the DCNN used in this paper, two neural networks, Visual Geometry Group (VGG)-11 and VGG-16, were selected for comparison experiments. In the experiments, VGG-11 adopted the standard network structure and VGG-16 adopted the improved network structure. The initial learning rate of both CNNs was 0.0001, and the rest of the parameters were the same as those of the DCNN. The experimental results are shown in Table 5.

Table 5. Classification results of different CNN[a] structures. The average accuracy rate is the percentage of all pulse type classifications that are correct. The average recall rate is the ratio of the correct pulse type in the classification result to the pulse type in the sample. Accuracy is the average of the accuracy of each of the 7 pulse types.

Classification model	Average accuracy rate (%)	Average recall rate (%)	Accuracy (%)
VGG[b]-11	74.4	74.7	74.7
VGG-16	77.3	77.3	77.4
DCNN[c]	79.1	78.9	79.1

[a]CNN: convolutional neural network.

[b]VGG: Visual Geometry Group.

[c]DCNN: deep convolutional neural network.

As can be seen from Table 5, the DCNN had the highest accuracy rate, which was 4.4%, and 1.7% higher compared with VGG-11 and VGG-16, respectively. In the average accuracy and average recall, the VGG-11 network had the lowest rates among the three models, which was 4.7% and 1.8% lower compared with the highest DCNN, respectively.

DSSN Experimental Results

To objectively evaluate the effectiveness of the model DSSN proposed in this paper, the base learners of the models were compared, and the experimental results are shown in Table 6.

Table 6. Classification results of different algorithms. The average accuracy rate is the percentage of all pulse type classifications that are correct. The average recall rate is the ratio of the correct pulse type in the classification result to the pulse type in the sample. Accuracy is the average of the accuracy of each of the 7 pulse types.

Classification model	Average accuracy rate (%)	Average recall rate (%)	Accuracy (%)
Time and frequency domain feature+SVM[a]	74.6	72.9	72.8
Time domain feature+SVM	79.2	76.2	76.1
DCNN[b]	79.1	78.9	79.1
DSSN[c]	83.2	82.9	83.3

[a]SVM: support vector machine.

[b]DCNN: deep convolutional neural network.

[c]DSSN: DCNN- and SVM-based stacking network.

As can be seen from Table 6, the DSSN model had the highest classification accuracy among the five methods, reaching 83.3%. In the average accuracy and average recall, the DSSN model rates improved by 8.6% and 10%, respectively, compared with the lowest time frequency domain feature model, and 4.1% and

4%, respectively, compared with the remaining highest DCNN model.

Discussion

Using machine learning or deep learning alone for pulse classification is not effective, but integrating both for learning can improve the classification accuracy of pulse signals. At the same time, existing research results on interpretable features of pulse signals can be absorbed and deep learning algorithms developed by technology can be used to further explore the information carried by pulse signals.

Comparison of Seven Pulse Type Classification Results Using an SVM

As can be seen from Figures 5 and 6, the recognition accuracy of the time domain classification model was higher than that of the time frequency domain classification model in the slippery, thready slippery, thready, thready stringy, stringy slippery, and stringy pulses, reaching 83%, 78%, 77%, 59%, 67%, and 76%, respectively, which was 2%-8% higher compared with the time frequency domain classification model. However, the time frequency domain classification model had the highest recognition rate of 95% for the flat pulse, which was 1% higher compared with the time domain classification model.

Figure 5. Confusion matrix of time domain features by SVM classification. SVM: support vector machine.

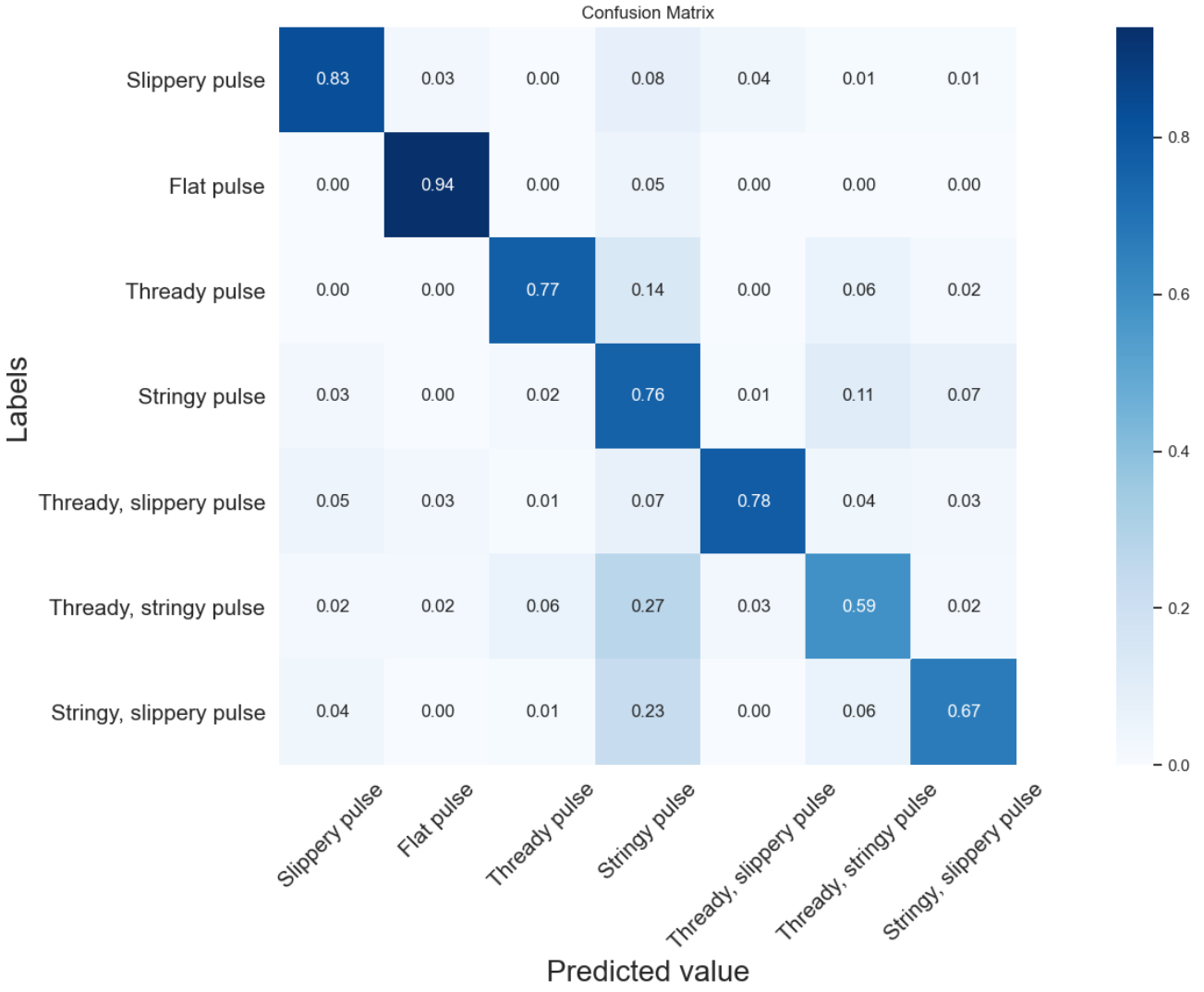
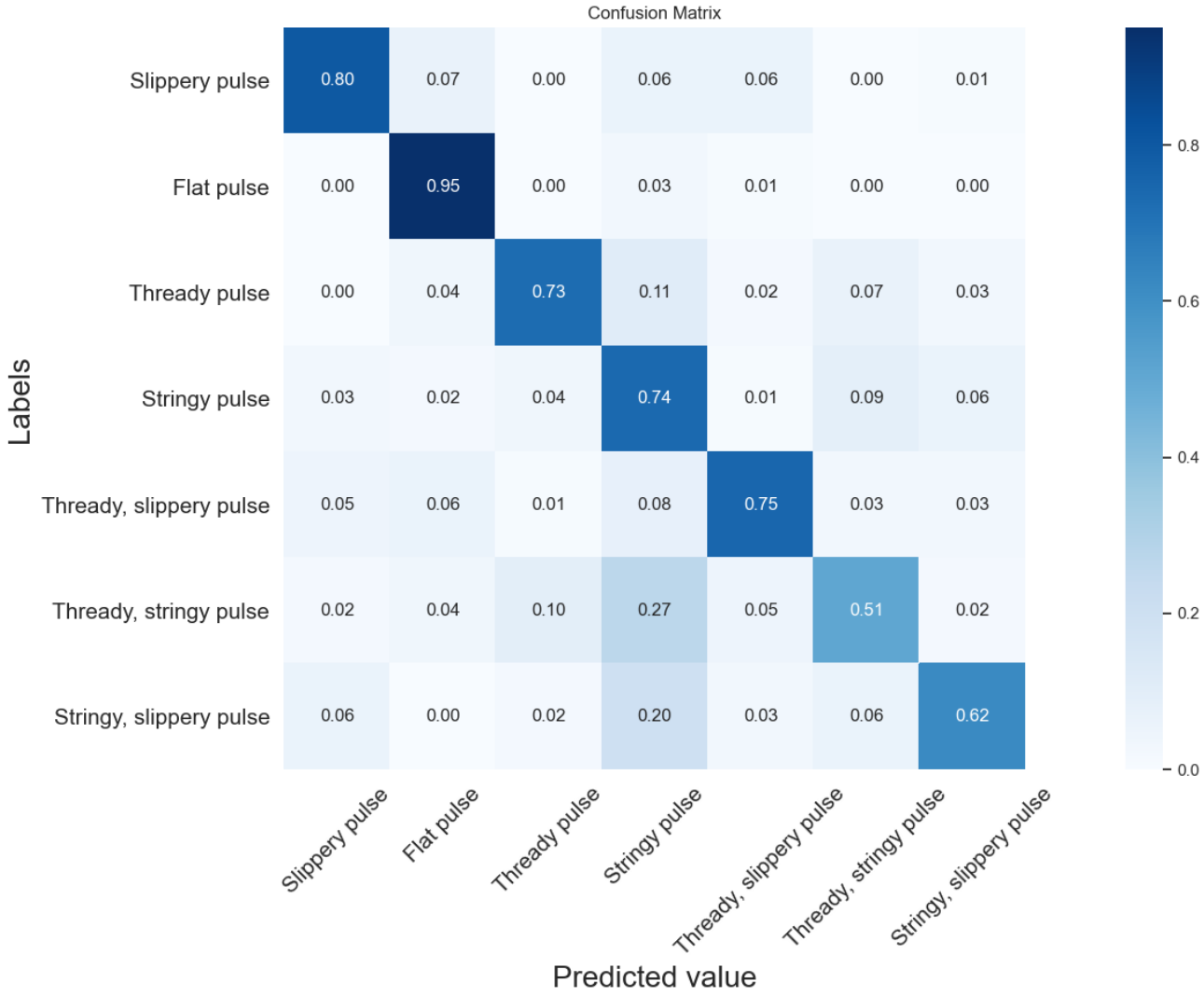


Figure 6. Confusion matrix of time and frequency domain features by SVM classification. SVM: support vector machine.



Overall, the time domain classification model was slightly better than the time frequency domain classification model. With the exception of the thready, stringy pulse, the accuracy of the recognition of the remaining pulse types was not significantly different. It can also be seen from the figure that the thready, stringy pulse and the stringy, slippery pulse were the pulse types with the lowest classification accuracy, and most of them were incorrectly classified as stringy pulses. In the time domain classification model, the recognition error rates were 27% and 23% for the thready, stringy and the stringy, slippery pulses, respectively; in the time frequency domain classification model, the recognition error rates were 27% and 20%, respectively. The reason for this situation might be that both the thready, stringy pulse and the stringy, slippery pulse have the characteristics of a stringy pulse, and it is difficult for the classifier to accurately determine their pulse types, leading to misclassification.

Because of the large number of time domain features extracted in this paper, only some features were selected for statistical analysis, as shown in Table 7. Among the seven types of pulse data, the main wave slope k of the slippery pulse was the largest, and the main wave amplitude h_1 was only second to the stringy, slippery pulse, which showed the characteristics of the high and steep main wave of the slippery pulse. The one-third pulse width w and pulse width period ratio w/T of the stringy pulse were the largest among the seven types of pulses, which showed the waveform characteristics of the wide main wave of the stringy pulse. The main wave amplitude h_1 of the stringy, slippery pulse was the largest, the main wave slope k was only lower than that of the slippery pulse, and the one-third pulse width w and pulse width period ratio w/T were larger, which showed that the stringy, slippery pulse had the waveform characteristics of both the stringy pulse and the slippery pulse, which was consistent with the characteristics of both pulses.

Table 7. Statistical analysis results of some time domain features. Feature parameters were h1, k, w, and w/T for 7 pulse types.

Feature parameters	Slippery pulse	Flat pulse	Thready, slippery pulse	Thready pulse	Thready, stringy pulse	Stringy, slippery pulse	Stringy pulse
h_1	610.701 ± 206.724	514.706 ± 121.698	356.757 ± 93.869	400.857 ± 139.692	407.703 ± 157.803	636.707 ± 227.226	592.404 ± 228.413
k	6.609 ± 2.241	5.886 ± 1.394	3.912 ± 1.002	3.662 ± 1.161	3.957 ± 1.468	6.329 ± 2.420	5.919 ± 2.543
w	0.126 ± 0.018	0.143 ± 0.031	0.140 ± 0.030	0.199 ± 0.034	0.205 ± 0.043	0.176 ± 0.036	0.207 ± 0.044
w/T	0.162 ± 0.023	0.165 ± 0.036	0.185 ± 0.040	0.214 ± 0.035	0.247 ± 0.038	0.220 ± 0.033	0.245 ± 0.038

The time domain features not only reflect the waveform characteristics of the pulse signal but also have certain physiological and pathological significance. The amplitude of the main wave, h_1, reflects the ejection function of the left ventricle and the compliance of the aorta; the amplitude of the main wave isthmus, h_2, has the same significance as the amplitude of the pre-repulse wave, h_3, and the sclerosis of blood vessels or the increase in peripheral resistance leads to an increase in the amplitudes h_2 and h_3. The amplitude of the descending isthmus, h_4, reflects the magnitude of peripheral resistance. The magnitude of the repulse wave, h_5, reflects the level of compliance of the aorta. The magnitude of phase t_1 reflects the rapidity of the left ventricular ejection time; the magnitudes of descending isthmus phase t_4 and repulse wave phase t_5 reflect the length of the systolic and diastolic phases of the left ventricle, respectively. The pulse period T indicates one cycle of the pulse, corresponding to one cardiac cycle of the left ventricle. The ratio of the descending isthmus main wave amplitude, h_4/h_1, reflects the level of peripheral resistance; the ratio of the repulse wave main wave amplitude, h_5/h_1, reflects the vascular compliance. The time ratio t_1/T reflects the rate of the cardiac ejection function, which increases when the rate decreases [18]. Wavelet packet analysis is used in the time frequency domain analysis to extract the energy magnitude of the pulse signal in different frequency bands, and its distribution reflects the elastic changes in the blood vessels [19], which also has some physiological significance.

In the classification algorithm, the SVM uses the kernel RBF to map the feature parameters into a high-dimensional space to provide better differentiability between different classes. The segmentation hyperplane is trained under this space to give the classification model the ability to recognize different pulse types. Thus, the time domain and time frequency domain features characterize the pulse signal from different perspectives, which can be combined with the SVM algorithm to obtain better classification results.

Comparison of Seven Pulse Type Classification Results Using VGG-11, VGG-16, and the DCNN

Figures 7-9 show confusion matrix plots of the three neural networks on the seven types of pulse data sets. As can be seen from the figures, the recognition accuracy of the DCNN improved in different degrees compared with VGG-11 and VGG-16 for the all categories of pulses except the flat pulse. Compared with VGG-11, the DCNN had a 6% higher recognition rate in the fine slippery pulse and the stringy, slippery pulse and a 5% higher recognition rate in the slippery pulse and the stringy pulse. Compared with VGG-16, the recognition rate was 1%-4% higher in all categories of pulses except the flat pulse. In the flat pulse recognition rate, VGG-16 had the highest accuracy of 98%, which was 3% higher than that of the DCNN. The recognition rate of the compound pulse was low overall, where most of the misclassified samples were classified as single-pulse types contained in the compound pulse, which might be due to the fact that the compound pulse had features that made up two of its single-pulse types, resulting in the classifier being unable to correctly identify its pulse type, similar to the results of the time-domain and time frequency-domain classification models. Overall, the DCNN had better classification performance compared with VGG-11 and VGG-16.

Figure 7. Confusion matrix of the seven types of pulses by VGG-11. The diagonal elements of the matrix indicate the prediction accuracy of different types of pulses. VGG: Visual Geometry Group.

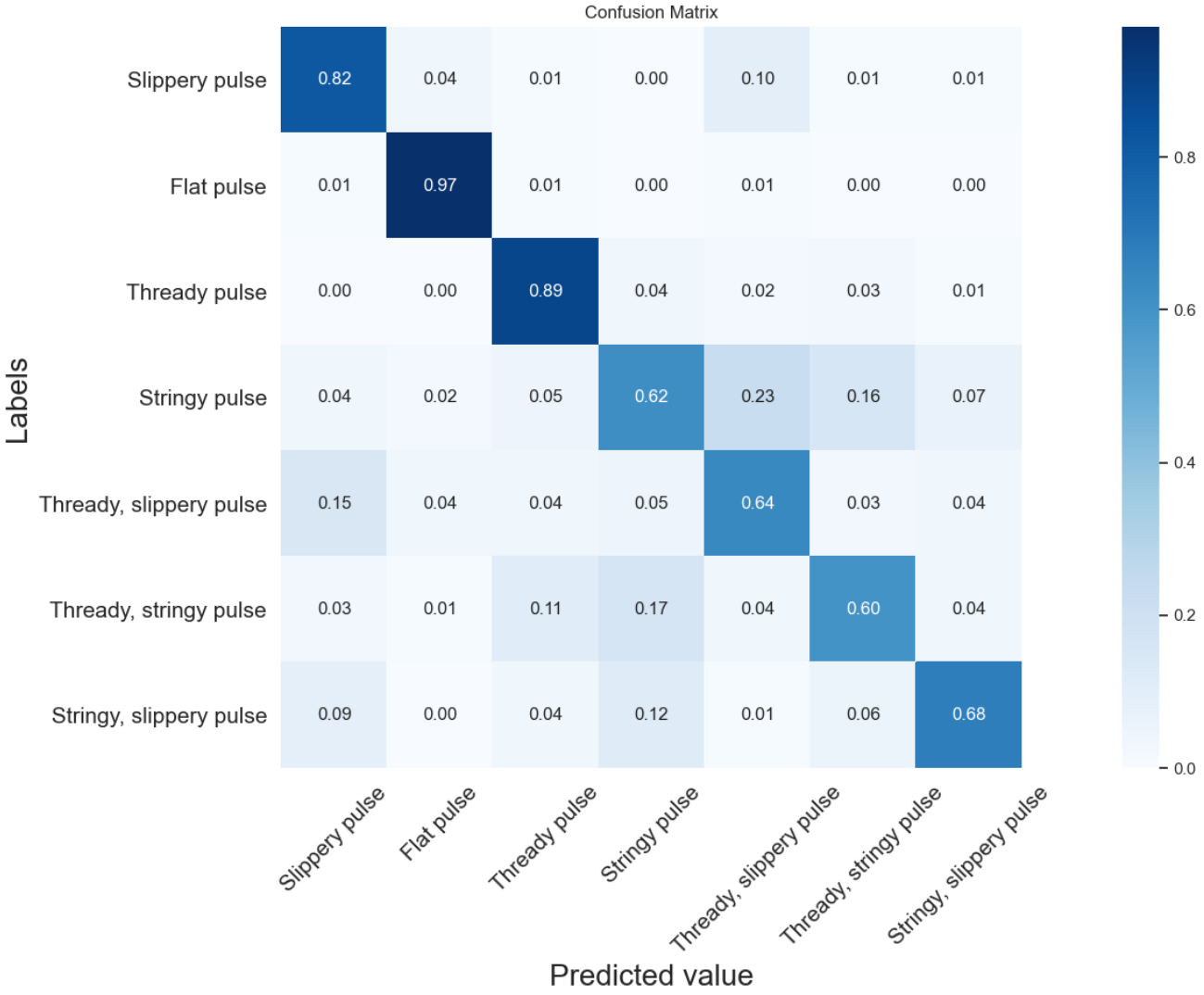
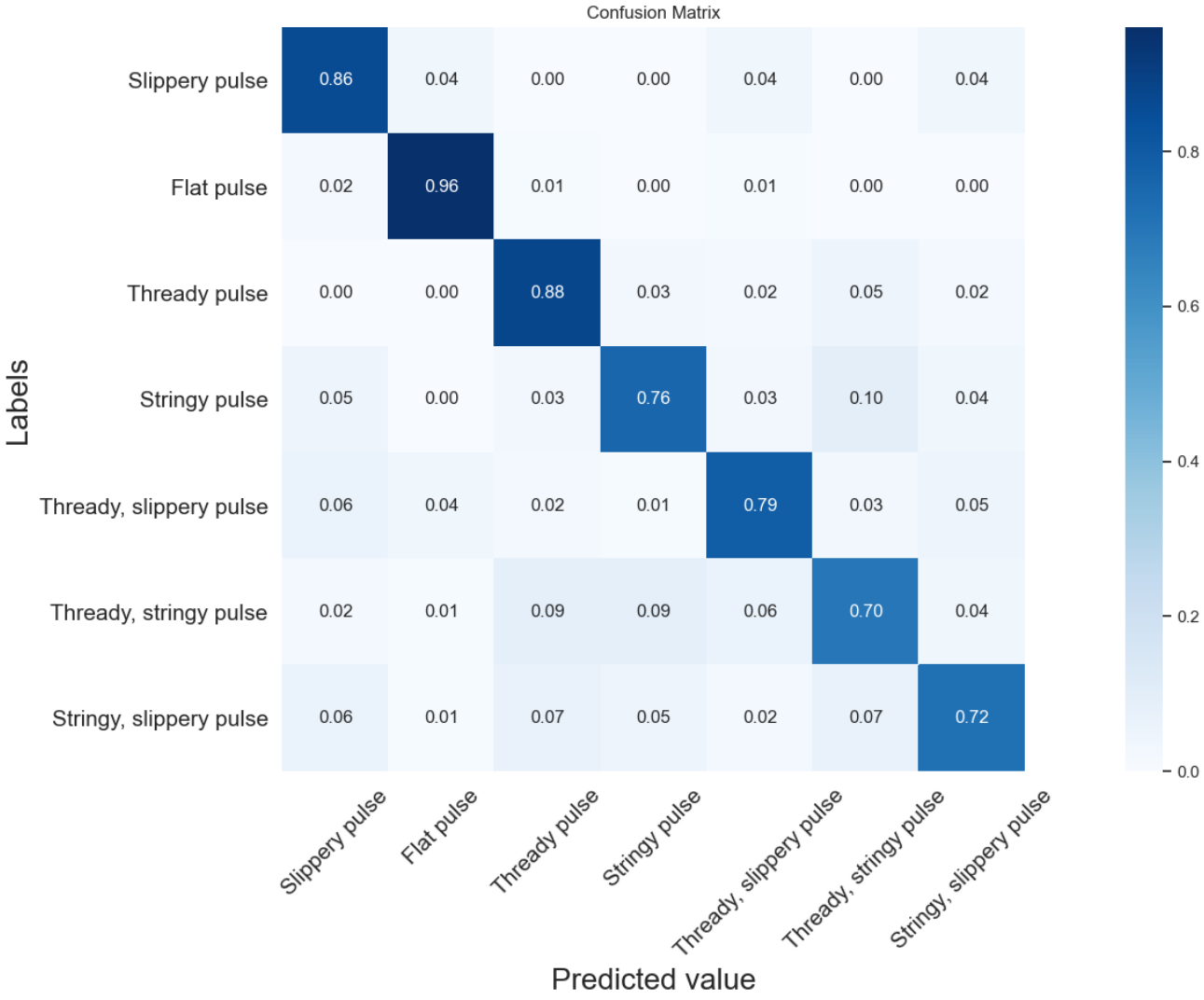


Figure 8. Confusion matrix of the seven types of pulses by VGG-16. The diagonal elements of the matrix indicate the prediction accuracy of different types of pulses. VGG: Visual Geometry Group.

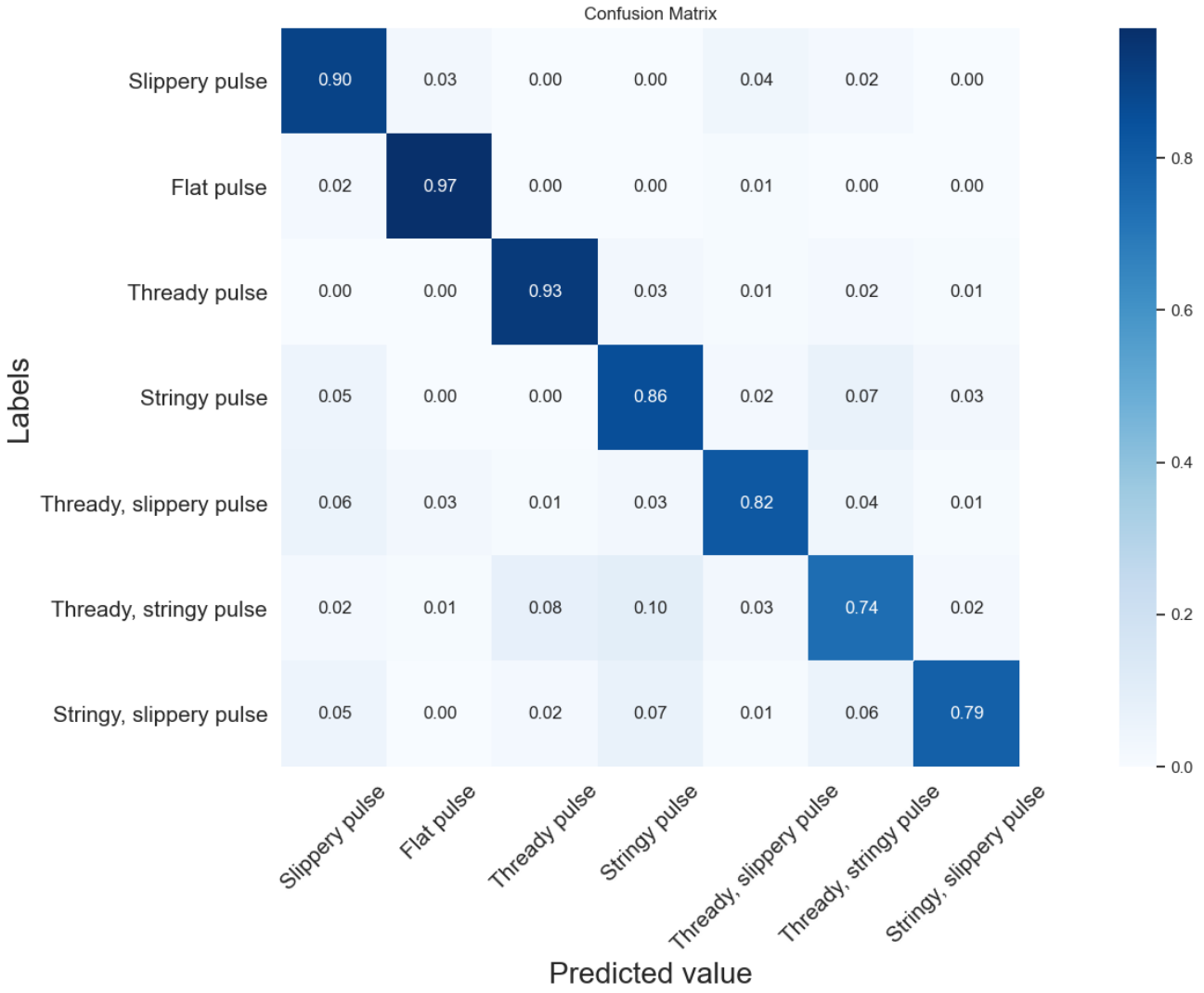
Figure 9. Confusion matrix of the seven types of pulses by the DCNN. The diagonal elements of the matrix indicate the prediction accuracy of different types of pulses. DCNN: deep convolutional neural network.



Comparison of Seven Pulse Type Classification Results Using a DSSN

It can be seen from Figure 10 that compared with the three base learners, the DSSN model improved the recognition rate in all seven pulse categories to varying degrees, with 3%, 2%, 4%, 1%, 10%, 5%, and 10% improvement compared to the highest recognition rate in each category of the base learners, respectively. Among them, the thready, stringy pulse, the stringy, slippery pulse, and the stringy pulse had the highest improvement effect, which reduced the recognition error rate to a greater extent. In addition, the DSSN model had higher recognition accuracy in each category than the DCNN model. Among them, the recognition rate of the stringy pulse was 10% higher compared with the DCNN model, and the remaining pulse types improved by 1%-7%. Therefore, the DSSN could integrate the advantages of multiple base learners and thus effectively improve the recognition accuracy of the model. Compared with existing pulse signal classification models, the DSSN also had better classification recognition results.

Figure 10. Confusion matrix of the DSSN. DSSN: deep convolutional neural network (DCNN)- and support vector machine (SVM)-based stacking network.



Although pulse types are clearly described in TCM textbooks, different TCM practitioners often interpret pulse diagnoses differently, depending on their own experience and understanding of pulses [20]. Even the same TCM practitioner may make different diagnoses for similar pulse characteristics in different circumstances. Using machine learning and deep learning methods for pulse classification can help TCM practitioners make better pulse diagnoses and improve the objectivity of the results. Using the DSSN method, the recognition accuracy for compound pulses can reach more than 70%, while for single pulses, the recognition accuracy is above 85%, and the best one can reach 97%. If the experimental samples can be enriched and the balance of samples can be improved, the accuracy of pulse classification can be further improved, which is promising for application in wearable devices.

Limitations and Conclusion

The existing work on pulse classification was mainly performed by machine learning or deep learning. In this paper, pulse classification was performed by a machine learning model (SVM) and a deep learning network (DCNN), but the results were not good. Through the DSSN method, the classification results of machine learning and deep learning were ensembled to obtain more accurate pulse type prediction results. Practitioners of TCM can use this method to assist in the diagnosis of TCM pulses, thus avoiding the uncertainty caused by subjectivity. Wearable devices can also use this method to determine the type of pulse of the user and thus predict the health status of the user, which is also relevant for the prevention of some diseases. At the same time, there were some areas for improvement in this experiment. First, the sample data of the pulse should be as large as possible, which will help improve the accuracy rate to some extent. Second, the diagnosis of the type of pulse signal collected should be integrated with the diagnosis results of several TCM experts, which can further improve the objectivity of the data.

A large number of original pulse wave studies have yielded many interpretable features, and many research results have been obtained for pulse wave classification. If only deep learning is used for classification, it will be difficult to use the results of previous research. Deep learning feature engineering and structured features reflect different pulse feature information that can complement each other. Therefore, based on the original pulse analysis, TCM scholars can combine the advantages of deep learning algorithms developed by technology to construct integrated classifiers that can provide better classification results

by making full use of the information obtained by deep learning feature engineering and artificially constructed features.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (nos. 82074332, 81673880, and 81302913), the Shanghai Science and Technology Committee Funding (no. 19441901100), and Shanghai Science and Technology Funding (no. 21DZ2271000).

Authors' Contributions

JY, XC, and SC designed this study. XC and SC performed data analysis. XC wrote the manuscript text, which was revised by JY. RG, HY, and YW helped with sample collection. All authors have read and approved the manuscript.

Conflicts of Interest

None declared.

References

1. O'Rourke M, Pauca A, Jiang X. Pulse wave analysis. Br J Clin Pharmacol 2001 Jun;51(6):507-522 [FREE Full text] [doi: 10.1046/j.0306-5251.2001.01400.x] [Medline: 11422010]
2. Korpas D, Hálek J, Dolezal L. Parameters describing the pulse wave. Physiol Res 2009 Dec;58(4):473-479 [FREE Full text] [doi: 10.33549/physiolres.931468] [Medline: 18656997]
3. Safar ME, Levy BI, Struijker-Boudier H. Current perspectives on arterial stiffness and pulse pressure in hypertension and cardiovascular diseases. Circulation 2003 Jun 10;107(22):2864-2869. [doi: 10.1161/01.CIR.0000069826.36125.B4] [Medline: 12796414]
4. Yamashina A, Tomiyama H, Arai T, Hirose KI, Koji Y, Hirayama Y, et al. Brachial-ankle pulse wave velocity as a marker of atherosclerotic vascular damage and cardiovascular risk. Hypertens Res 2003 Aug;26(8):615-622 [FREE Full text] [doi: 10.1291/hypres.26.615] [Medline: 14567500]
5. Cohn J, Finkelstein S, McVeigh G, Morgan D, LeMay L, Robinson J, et al. Noninvasive pulse wave analysis for the early detection of vascular disease. Hypertension 1995 Sep;26(3):503-508. [doi: 10.1161/01.HYP.26.3.503] [Medline: 17049917]
6. Anson CYT. Review of traditional Chinese medicine pulse diagnosis quantification. In: Complement Ther Contemp Healthc. Sao Paulo: Marcelo Saad; 2012:61-80.
7. Lee JY, Jang M, Shin SH. Study on the depth, rate, shape, and strength of pulse with cardiovascular simulator. Evid Based Complement Alternat Med 2017 Dec;2017(11-12):2867191-2867800 [FREE Full text] [doi: 10.1155/2017/2867191] [Medline: 28246538]
8. Xu LS, Meng MQ, Wang KQ. Pulse image recognition using fuzzy neural network. Annu Int Conf IEEE Eng Med Biol Soc 2007 Dec;2007(11-12):3148-3151. [doi: 10.1109/IEMBS.2007.4352997] [Medline: 18002663]
9. Xu L, Zhang D, Wang K, Wang L. Arrhythmic pulses detection using Lempel-Ziv complexity analysis. EURASIP J Adv Signal Process 2006 Mar 26;2006(1):1-12. [doi: 10.1155/asp/2006/18268]
10. Zhang D, Zuo W, Zhang D, Zhang H, Li N. Classification of pulse waveforms using edit distance with real penalty. EURASIP J Adv Signal Process 2010 Aug 30;2010(1):1-8. [doi: 10.1155/2010/303140]
11. Garmaev BZ, Boronoev VV, Naguslaeva IV, Ompokov VD. Classification of pulse waves based on cluster analysis of time parameters. J Phys: Conf Ser 2019 May 04;1210:012048-012048. [doi: 10.1088/1742-6596/1210/1/012048]
12. Li G, Watanabe K, Anzai H, Song X, Qiao A, Ohta M. Pulse-wave-pattern classification with a convolutional neural network. Sci Rep 2019 Oct 17;9(1):14930-14800 [FREE Full text] [doi: 10.1038/s41598-019-51334-2] [Medline: 31624300]
13. Huang CH, Wang YM, Smith S. Using high-dimensional features for high-accuracy pulse diagnosis. Math Biosci Eng 2020 Oct 09;17(6):6775-6790 [FREE Full text] [doi: 10.3934/mbe.2020353] [Medline: 33378877]
14. Fernandez A, Garcia S, Herrera F, Chawla N. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. jair 2018 Apr 20;61:863-905. [doi: 10.1613/jair.1.11192]
15. Chen YH, Yang CC, Cao QF, Li BT, Shang YS. The comparison of some time-frequency analysis methods. Prog Geophys 2006 Dec;21(4):1180-1185. [doi: 10.1016/S1001-8042(06)60011-0]
16. Guo Q, Wang K, Zhang D. A wavelet packet based pulse waveform analysis for cholecystitis and nephrotic syndrome diagnosis. 2008 Presented at: International Conference on Wavelet Analysis and Pattern Recognition; 2008; Hong Kong, China p. 30-31. [doi: 10.1109/icwapr.2008.4635834]
17. Deepa R, Rajaguru H, Ganesh Babu C. Analysis on wavelet feature and Softmax discriminant classifier for the detection of epilepsy. 2020 Dec Presented at: IOP Conference Series: Materials Science and Engineering; 2020.12.11; Tamil Nadu, India p. 012036-012038. [doi: 10.1088/1757-899x/1084/1/012036]

18. Zhang W, Zhang Y, Zhang S. Application on wavelet and neural network in the analysis and pattern recognition of the manifestation of the pulse for detection of cerebrovascular disease. Shanghai Biomed Eng 2008;29(2):84-86. [doi: 10.3969/j.issn.1674-1242.2008.02.006]
19. Hu X, Zhu H, Xu J. Wrist pulse signals analysis based on deep convolutional neural networks. 2014 Presented at: IEEE Conference on Computational Intelligence in BioinformaticsComputational Biology; 2014; Honolulu, HI, USA p. 21-24. [doi: 10.1109/cibcb.2014.6845525]
20. Alice LYL, Guan B, Chen C, Chan H, Kong K, Li W, et al. Artificial intelligence meets traditional Chinese medicine: a bridge to opening the magic box of sphygmopalpation for pulse pattern recognition. Digital Chin Med 2021 Mar;4(1):1-8 [FREE Full text] [doi: 10.1016/j.dcmed.2021.03.001]

Abbreviations

CNN: convolutional neural network
CVD: cardiovascular disease
DCNN: deep convolutional neural network
DSSN: DCNN- and SVM-based stacking network
ERP: edit distance with real penalty
FCNN: fully connected neural network
KNN: k-nearest neighbor
RBF: radial basis function
SGD: stochastic gradient descent
SMOTE: synthetic minority oversampling technique
SVM: support vector machine
TCM: traditional Chinese medicine
VGG: Visual Geometry Group

Edited by C Lovis; submitted 18.02.21; peer-reviewed by L Wang, D Oladele, D Hu; comments to author 10.07.21; revised version received 12.08.21; accepted 25.09.21; published 21.10.21.

Please cite as:
Yan J, Cai X, Chen S, Guo R, Yan H, Wang Y
Ensemble Learning-Based Pulse Signal Recognition: Classification Model Development Study
JMIR Med Inform 2021;9(10):e28039
URL: https://medinform.jmir.org/2021/10/e28039
doi:10.2196/28039
PMID:34673537

Original Paper

Adverse Drug Event Prediction Using Noisy Literature-Derived Knowledge Graphs: Algorithm Development and Validation

Soham Dasgupta[1], HS; Aishwarya Jayagopal[2], BTech; Abel Lim Jun Hong[2], BComp; Ragunathan Mariappan[2], MSc; Vaibhav Rajan[3], PhD

[1]Mallya Aditi International School, Bangalore, India

[2]School of Computing, National University of Singapore, Singapore, Singapore

[3]Department of Information Systems and Analytics, National University of Singapore, Singapore, Singapore

Corresponding Author:
Vaibhav Rajan, PhD
Department of Information Systems and Analytics
National University of Singapore
School of Computing
Singapore, 117417
Singapore
Phone: 65 65166737
Email: vaibhav.rajan@nus.edu.sg

Abstract

Background: Adverse drug events (ADEs) are unintended side effects of drugs that cause substantial clinical and economic burdens globally. Not all ADEs are discovered during clinical trials; therefore, postmarketing surveillance, called pharmacovigilance, is routinely conducted to find unknown ADEs. A wealth of information, which facilitates ADE discovery, lies in the growing body of biomedical literature. Knowledge graphs (KGs) encode information from the literature, where the vertices and the edges represent clinical concepts and their relations, respectively. The scale and unstructured form of the literature necessitates the use of natural language processing (NLP) to automatically create such KGs. Previous studies have demonstrated the utility of such literature-derived KGs in ADE prediction. Through unsupervised learning of the representations (features) of clinical concepts from the KG, which are used in machine learning models, state-of-the-art results for ADE prediction were obtained on benchmark data sets.

Objective: Due to the use of NLP to infer literature-derived KGs, there is *noise* in the form of false positive (erroneous) and false negative (absent) nodes and edges. Previous representation learning methods do not account for such inaccuracies in the graph. NLP algorithms can quantify the confidence in their inference of extracted concepts and relations from the literature. Our hypothesis, which motivates this work, is that by using such confidence scores during representation learning, the learned embeddings would yield better features for ADE prediction models.

Methods: We developed methods to use these confidence scores on two well-known representation learning methods—DeepWalk and Translating Embeddings for Modeling Multi-relational Data (TransE)—to develop their *weighted* versions: Weighted DeepWalk and Weighted TransE. These methods were used to learn representations from a large literature-derived KG, the Semantic MEDLINE Database, which contains more than 93 million clinical relations. They were compared with Embedding of Semantic Predications, which, to our knowledge, is the best reported representation learning method using the Semantic MEDLINE Database with state-of-the-art results for ADE prediction. Representations learned from different methods were used (separately) as features of drugs and diseases to build classification models for ADE prediction using benchmark data sets. The methods were compared rigorously over multiple cross-validation settings.

Results: The *weighted* versions we designed were able to learn representations that yielded more accurate predictive models than the corresponding unweighted versions of both DeepWalk and TransE, as well as Embedding of Semantic Predications, in our experiments. There were performance improvements of up to 5.75% in the F_1-score and 8.4% in the area under the receiver operating characteristic curve value, thus advancing the state of the art in ADE prediction from literature-derived KGs.

Conclusions: Our classification models can be used to aid pharmacovigilance teams in detecting potentially new ADEs. Our experiments demonstrate the importance of modeling inaccuracies in the inferred KGs for representation learning.

(JMIR Med Inform 2021;9(10):e32730) doi:10.2196/32730

KEYWORDS

adverse drug event; knowledge graph; Embedding of Semantic Predications; biomedical literature

Introduction

The Challenge of Detecting Adverse Drug Events

Adverse drug events (ADEs) are unintended side effects of drugs that often lead to emergency visits, prolonged hospital stays, and worse patient outcomes [1]. They pose substantial clinical and economic burden—in the United States alone, morbidity and mortality costs associated with ADEs were estimated to be approximately US $528 billion in 2016 [2], and 1 in 3 drugs approved in the period from 2000 to 2010 had safety-related issues after release, some of which led to their withdrawal from the market [3].

Patients may be prescribed multiple drugs together when they have multiple coexisting ailments or for combination therapies, for example, in cancer [4]. In such cases, it is also possible for ADEs to occur because of a combination of drugs, also termed polypharmacy. Polypharmacy poses a higher risk of ADEs because of drug-drug interactions [5,6]. Polypharmacy is also an increasing burden to health care; estimates suggest that they cause nearly 74,000 emergency room visits and 195,000 hospitalizations annually in the United States [7].

In general, detecting ADEs is a challenging problem. Clinical trials are limited by the number and characteristics of patients tested as well as the duration of the observation period, and they may not detect all ADEs, especially those with long latency or those that affect only certain patient groups [8]. Detecting polypharmacy ADEs is even harder—although it is possible to test for a few drug interactions [9], it is computationally infeasible to test for all possible drug combinations [10]. Postmarketing drug safety surveillance, called pharmacovigilance, is routinely conducted to continuously update our knowledge of potential ADEs.

Spontaneous reporting systems, which collect voluntary reports of ADEs, have been the primary data source for pharmacovigilance. Mining these databases presents several challenges because of inherent reporting bias and incompleteness. Methods to detect ADE signals from other data sources such as social media and clinical data are being actively developed, but problems of quality and reliability limit the utility of these sources; the study by Ventola [1] provides a detailed survey. Biomedical literature, which forms another source of ADE signals, is also consulted during ADE mining from other sources. An advantage of these data over others is the presence of information relevant to potential causal assessment in the studies described. Furthermore, as biomedical knowledge grows, this source continues to expand and update itself systematically.

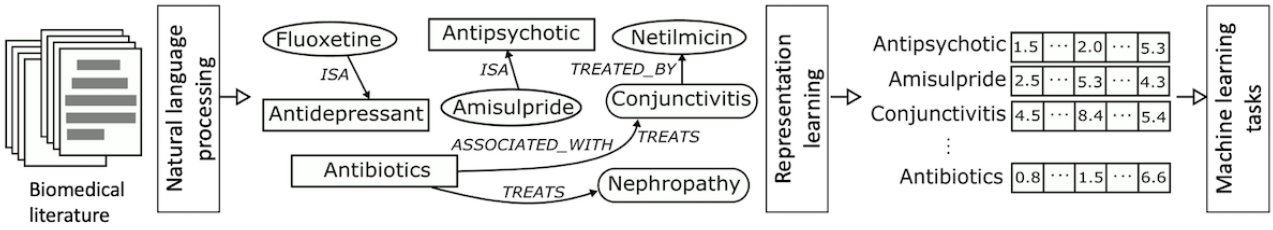
However, the scale is both an advantage and a hurdle. MEDLINE, the largest index of medical literature, contains more than 24 million articles, with more than a million new articles published annually [11]. This enormous scale makes it challenging to mine the data; therefore, to facilitate knowledge discovery from such unstructured data, standardized vocabularies and ontologies have been created. Furthermore, natural language processing (NLP) techniques have been developed to automatically infer both clinical concepts and their relations found in the literature. Such ontologies are also massive and continue to evolve with growing biomedical literature. An ontology can be viewed as a heterogeneous knowledge graph (KG) comprising multiple kinds of vertices (clinical concepts, eg, drugs and diseases) and edges (relations, eg, Treats and Is-A-Side-Effect).

Previous literature-based knowledge discovery systems—for ADE detection as well as for other applications—have mainly used text and graph mining or supervised learning methods [12-14] that require careful design of the features from text or graphs. For instance, certain patterns of relations (edges) among clinical concepts (vertices) may be used to mine ontologies for a potential ADE. Finding the right set of patterns can be challenging—for a given pair of concepts, evidence of an association, or lack thereof, cannot be discerned from the presence or absence of a single edge: two clinical concepts may be indirectly connected and, by multiple paths, be composed of several relations. Such manual feature engineering is cumbersome, time consuming, and does not scale with the rapidly evolving literature and literature-derived KGs.

To enable reasoning on such large and complex KGs and to automate feature engineering, most recent approaches use *graph embeddings* that encode the global structural properties of a given graph into vectorial *representations* of its vertices. With such representations, relations among clinical concepts can be computed algebraically using vectorial measures of similarity. Furthermore, these representations can be used as features directly in machine learning models for tasks such as association prediction or cluster detection (see Figure 1 for a schematic). Such approaches have yielded state-of-the-art results in many tasks, including ADE prediction from KGs [15].

Most representation learning methods have been designed for graphs from the internet, for example, social media or e-commerce, where the graph itself is assumed to have very few or no errors. In contrast, errors are common in literature-derived biomedical KGs because of the large and complex clinical vocabulary, which often contains inconsistently used abbreviations and features frequent use of synonyms and homonyms, as well as the need to use and link multiple expert-curated ontologies and insufficient labeled data sets for the underlying NLP tools used [16-18]. The best previous embeddings, Embedding of Semantic Predications (ESP; detailed in the section *ESP Method*), which was designed for such literature-derived graphs, does not account for such *noise* due to NLP inference.

Figure 1. Knowledge graphs are obtained through natural language processing on biomedical literature. Clinical concept representations learnt from such graphs are used as features for machine learning tasks. ISA: is a.



NLP algorithms quantify the confidence in their inference of extracted concepts and relations from the literature. Our hypothesis, which motivated this work, is that by using such confidence scores during representation learning, the learned embeddings would yield better features for predictive models. In this study, we developed techniques to use these confidence scores during representation learning to model inaccuracies in literature-derived KGs due to NLP inference. We illustrate the use of our technique on two well-known representation learning methods: DeepWalk [19] and Translating Embeddings for Modeling Multi-relational Data (TransE) [20]. We show how confidence scores can easily be incorporated in both these methods as *weights* that bias the methods to choose higher confidence edges and nodes over lower confidence edges and nodes during representation learning. Thus, we developed the *weighted* versions of these methods: *Weighted DeepWalk* and *Weighted TransE*.

We rigorously evaluated these methods on benchmark data sets for drug-ADE prediction and polypharmacy prediction. In both tasks, our weighted versions were able to learn representations that yielded more accurate predictive models than ESP and the unweighted versions of DeepWalk and TransE, with improvements of up to 5.75% in the F_1-score and 8.4% in the area under the receiver operating characteristic curve (AUC) value. Thus, our experimental results demonstrate the benefit of modeling inaccuracies in the inferred KGs for representation learning. Better representations, in turn, lead to better classification models for ADE prediction.

Background and Related Work

Biomedical KGs

The primary source of scientific clinical knowledge is biomedical literature, which records details of clinical trials conducted, case studies, and systematic reviews. To facilitate knowledge discovery from such unstructured data, standardized vocabularies and ontologies have been created; for example, the Unified Medical Language System Metathesaurus [21] contains more than 5 million clinical concepts—identified by controlled unique identifiers—that have been organized into structured ontologies.

NLP techniques that have been designed to infer both clinical concepts and their relations found in the literature automatically create ontologies from the rapidly growing body of biomedical literature. Data sources such as molecular databases, drug banks, or social media may also be used as additional inputs. Examples include the Semantic MEDLINE Database (SemMedDB) [22] and KnowLife [23]. These automatically generated ontologies have been found to be immensely useful to support hypothesis generation [12], literature-based knowledge discovery [24], and predictive modeling [25].

In this work, we used the SemMedDB, where clinical concepts are identified in PubMed abstracts through entity recognition algorithms and then mapped to their controlled unique identifiers. Various heuristics are used to infer the relations between concepts (see the study by Rindflesch and Fiszman [18] for details). The SemMedDB infers 30 different kinds of relations such as *Treats*, *Causes*, *Predisposes*, and *Prevents* among clinical concepts of various types that include diseases, drugs, procedures, and biological structures. These relations are organized into [*subject-predicate-object*] triplets (eg, [drug A-Treats-disease B]), where both the subject and object are clinical concepts and the predicate is a relation. There are more than 96 million such triplets extracted in the SemMedDB.

The SemMedDB contains useful information about each triplet including the following:

1. Co-occurrence scores of [subject-predicate-object] triplet: the number of times the triplet is inferred from the literature—higher number indicates higher confidence in the association.

(2) Subject-score and object-score: confidence score of the mapping found by NLP recognition algorithms between a text string and the subject or object concept. These scores were used in the methods we developed.

The collection of such triplets can be viewed as a *heterogeneous graph* comprising multiple vertex types (clinical concepts) and multiple edge types (predicates). This graph-based view enabled us to mine the KG using graph analytics tools. For instance, predicates only show direct relations between two concepts, whereas the graph illuminates indirect relations through various paths connecting the two concepts.

Learning Clinical Concept Representations From KGs

Graph Embeddings

Statistical machine learning models typically assume inputs as feature vectors. To obviate the need for manual extraction of features from text and graph inputs, representation learning aims to learn features or representations from the input directly, in an unsupervised manner. Representation learning from graphs is an active research area; see the studies by Goyal and Ferrara [26] and Yang et al [27] for general surveys and the study by Wang et al [28] for a survey on representation learning on KGs. The representations are vectorial representations of the vertices of the graph. They are also called graph *embeddings* because

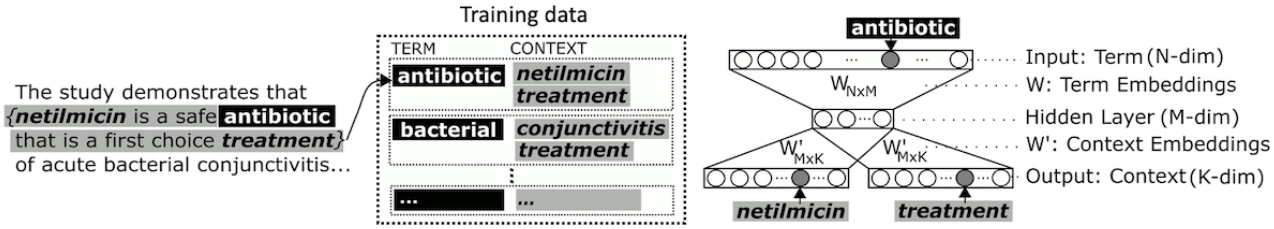
of the similarities in both the representation learning algorithms and their uses to the word embeddings in NLP.

Formally, for any given graph $G = (V, E)$, with vertex set V and edge set E, a graph representation learning algorithm learns a d-dimensional latent representation $x_{vi} \in \mathbb{R}^d$, $d \ll |V|$ for each vertex $v_i \in V$ that captures global structural and semantic relations (as described below) in the graph. We first outline skip-gram negative sampling (SGNS), a widely used neural architecture for obtaining word embeddings in NLP, and later describe how SGNS is used or adapted to obtain graph embeddings.

SGNS Approach

SGNS is a neural approach to learn word representations from text data [29]. The key idea is to train a neural network to predict the *context* of each word in the input text corpus, where context is defined as a window of neighboring words. Usually, preprocessing steps remove uninformative words such as stop-words (*a*, *an*, *the*...) before training, and one-hot encoding is used for input and output of the network. The window size is a parameter set during training. For each word, context words for every occurrence in the text corpus are extracted to form the training data (Figure 2). After the network is trained using gradient descent, the learned weights are used as word embeddings. The model can use negative samples—where words not found in the neighborhood are used—during training.

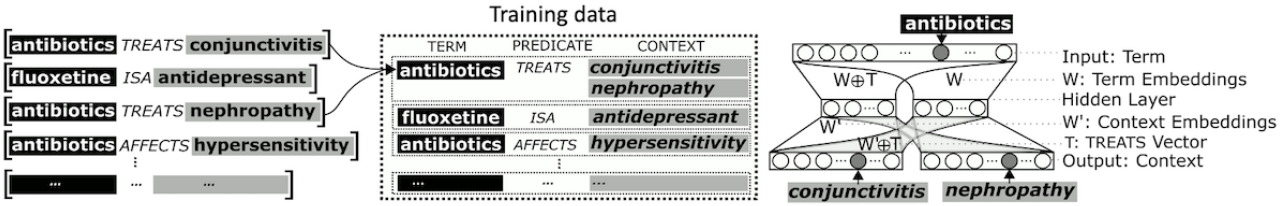
Figure 2. Skip-gram negative sampling: A window of words around a term constitutes its context. Word embeddings are obtained from the weights of a neural network trained to predict context words of a term. K-dim: K dimensions; M-dim: M dimensions; N-dim: N dimensions; SGNS: skip-gram negative sampling.



ESP Method

To obtain embeddings of clinical concepts using [subject-predicate-object] triplets, also called predications, Cohen and Widdows [30] designed the ESP method based on the SGNS architecture. In ESP, the context of a *subject* concept is defined as the set of *objects* that it relates to through one or more predicates. In addition, their model was explicitly trained to enable analogical reasoning, with respect to biomedical relations, by defining binding operators (eg, exclusive OR [XOR], denoted by ⊕) on the representations of concepts and predicates. Thus, if there is a predicate such as drug A-Treats-disease B, then from the corresponding representations, they aim to have *drug A⊕treats≈disease B*. This is learned during training by modifying the SGNS architecture to predict the object (eg, disease B) from the XOR of the predicate and subject (drug A⊕treats; Figure 3). With these modifications, ESP obtains embeddings of both clinical concepts and predicates using a gradient descent–based optimization similar to that of SGNS.

Figure 3. Embedding of Semantic Predications embeddings from [subject-predicate-object] triplets: the objects of a subject term form its context, and the skip-gram negative sampling architecture is modified to predict each context from the term and the predicate. ISA: is a.



ESP has not been developed by viewing the collection of triplets as a KG. When viewed from the KG perspective, we recognize that ESP trains its embeddings by using only its immediate neighbors in the graph. It is possible to learn embeddings that can explicitly incorporate more distant information, that is, by using a context for training that includes not just neighboring vertices but also vertices that are two or more hops away on the KG, for instance, through walk-based approaches that we describe next.

DeepWalk Method

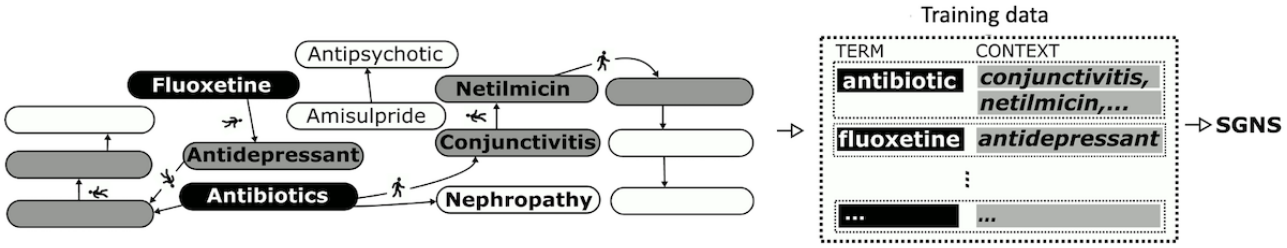
There are many graph embedding algorithms based on random walks. Although the details differ, they share the underlying idea of using random walks on the graph to define a context for a vertex and to generate training data similar to those for learning word embeddings. Then a neural architecture can be used to obtain vertex representations. We outline DeepWalk [19], one such walk-based algorithm. DeepWalk obtains training data through random walks from each vertex on the input graph and uses SGNS to obtain vertex representations (Figure 4). Note that DeepWalk assumes a homogeneous input graph; information regarding multiple vertex types and edge labels is not used and, hence, is not shown in Figure 4. The random walk generator randomly selects the next vertex to walk to from its neighborhood, that is, vertices that are connected by an edge. For each vertex in the input graph, select N sequences of L

vertices each. In each sequence, at the kth step, the k+1st vertex is randomly sampled with probability:

where $H(v_j)$ denotes the neighborhood of v_j. Additional implementation details can be found in the study by Perozzi et al [19].

The simple idea of DeepWalk has been extended in many ways for homogeneous graphs. Relatively fewer walk-based approaches have been developed for heterogeneous graphs. Among them, Metapath2Vec is an effective method that uses *metapath schemes* to predefine the types of edges to be selected during random walk selection [31], an approach that works well in sparse graphs with relatively few edge types. However, the generation of such metapath schemes is difficult for biomedical KGs that are typically very dense and have many edge types. There are approaches, which are different from walk-based approaches, that have been designed directly for KGs, such as TransE, which we describe next.

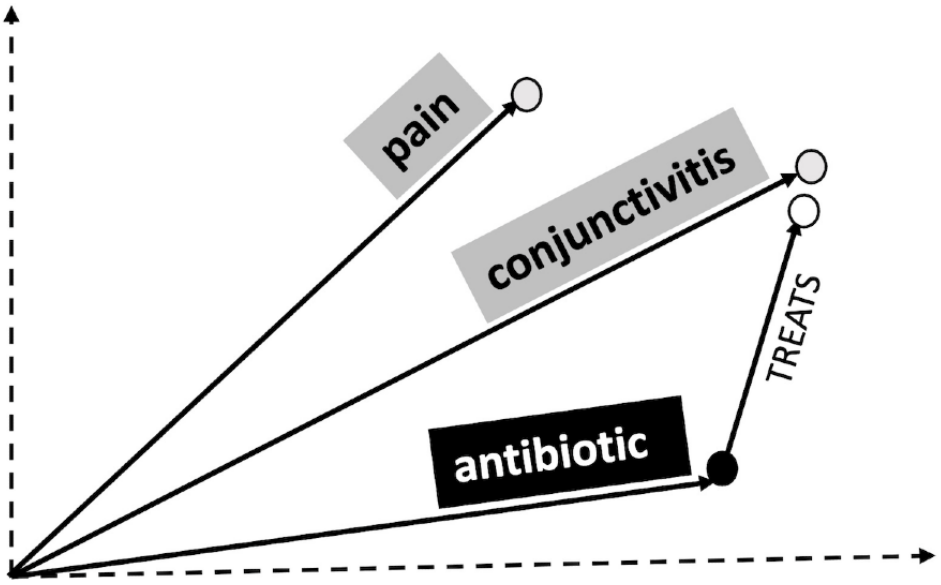
Figure 4. DeepWalk: Random walks generate contexts for a vertex, which are used as training data in skip-gram negative sampling to obtain embeddings for each vertex. SGNS: skip-gram negative sampling.



TransE Method

We briefly describe the intuition behind TransE and refer the reader to the study by Bordes et al [20] for more details. For a set of clinical entities E, given a training set S of triplets (h, l, t)—composed of two clinical entities head h and tail t where h, $t \in$ E—and a relationship (or predication) l, the TransE model learns k-dimensional vector representations of the entities and the relationships (where k is a hyperparameter). The idea behind the TransE model is that the relationship induced by the l-labeled edges corresponds to a translation of the vector representations. That is, we want that the vectors obey $h + l \sim t$ when the predicate (h, l, t) is present in the KG. If the triplet (h, l, t) is not present, then the vector $h + l$ should be far away from the tail concept t in vector space (Figure 5).

Figure 5. Schematic of TransE: Triplet (antibiotic, treats, conjunctivitis) is preserved in the vector sum of their representations in 2 dimensions: h[antibiotic] + l[treats] ~ t[conjunctivitis]. The vector sum h + l is much further from the vector for pain than from the vector for conjunctivitis. TransE: Translating Embeddings for Modeling Multi-relational Data.



Note that the input of TransE is similar to that of ESP: both use [subject-predicate-object] triplets. However, ESP generates binary representations and has a different scheme for the composition of representations, whereas TransE obtains real-valued distributed representations with the usual operations defined on the vector space. Furthermore, TransE does not use SGNS for training and has a different energy-based framework for optimization as described below.

Following an energy-based framework, the energy of a triplet is given by $d(h + l, t)$, where d is a dissimilarity function (eg, L1 or L2 norm); lower energy triplets are preferred because they preserve the required vector relationship. Therefore, to learn the vector representations, a margin-based ranking

criterion, given below, is minimized over the set of triplets in the KG:

where $[x]_+$ denotes the positive part of x and $y>0$ is a margin hyperparameter and $S'_{(h,l,t)} = \{(h', l, t)|h' \in E\} \cup \{(h, l, t')|t' \in E\}$ is the set of corrupted triplets. The set S' is composed of training triplets with either the head diseases or tail diseases replaced by a random entity (but not both at the same time). The loss function L favors lower values of the energy for training triplets than for corrupted triplets and is thus a natural implementation of the intended criterion. The model uses stochastic gradient descent optimization to minimize the loss.

Methods

Model 1: Extending DeepWalk to Weighted DeepWalk

We made two modifications to DeepWalk. The first modification enabled us to sample edges in a heterogeneous graph without requiring fixed predefined metapath schemes. We then introduced a bias over the walks that was informed by simple statistics of the inferred clinical concepts and relations and, thus, accounted for inaccuracies during NLP inference. Both these approaches change the sampling strategy in the procedure for generating random walks of DeepWalk. Other steps involving SGNS for training remain the same.

We modified the random walk procedure such that an (edge, vertex) pair was selected for traversal at each step instead of just a vertex. Thus, if the same vertex could be reached through two different predicates (edges), they were considered two separate neighbors during the next step selection in the random walk. We viewed a subject vertex as one that was connected to not just another vertex, but also to a pair (predicate, object). Formally, we defined the set $E' = \{((p, o), s)$ iff $[s, p, o]$ is a valid triplet$\}$. This defines the neighborhood of a vertex v_j as a set of (predicate, vertex) pairs: $H'(v_j) = \{(p, v_i)|((p, v_i), v_j) \in E'\}$. We incorporated information on the confidence scores from the SemMedDB using a scoring function in the sampling distribution at each step of the walk. As a result, the walks were biased toward vertices and edges with higher confidence. The selection of the next (edge, vertex) pair was performed by sampling from the distribution:

where f_{ijp} is a score for the corresponding triplet and $\sigma_{N'(vj)}$ is a softmax function over all the predicates from vertex v_j. The triplet score was computed as a weighted product,  where $W_{ijp} = (w_j s_{vj} \times w_i s_{vi} \times w_p c_p)$ and s_v represents the score for the (subject and object) vertices and c_p represents the co-occurrence score of the predicate.

The normalization, using the maximum value, was carried out to avoid numerical errors due to very large numbers. Each score had a multiplicative effect that resulted in triplets, with all 3 high scores being highly favored (for the next vertex selection in a walk) over triplets with any of the 3 scores being low. The weights were optimized through hyperparameter search. The softmax function was used to convert the scores to probabilities at each step of the walk. We performed L2 normalization on the learned representations to ensure that their L2 norms were equal to 1. We implemented both the random walk generators in Python (Python Software Foundation) and used the script from the study by Mikolov et al [29] for SGNS.

Model 2: Extending TransE to Weighted TransE

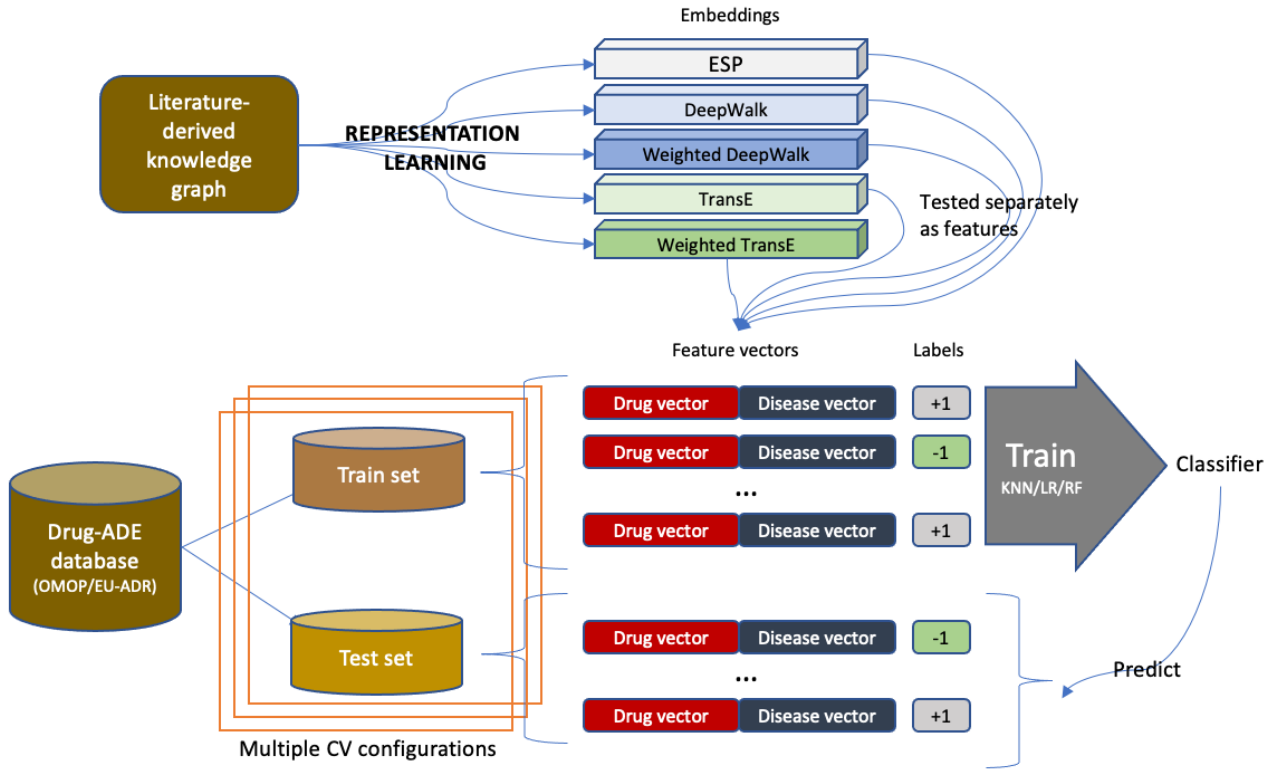
Incorporating confidence scores in TransE is relatively straightforward. As described earlier, the loss function is designed in such a way that true triplets have lower energy than corrupted triplets. Using the weight function of the subject, object, and co-occurrence scores—f_{ijp}, defined earlier for Weighted DeepWalk—we can simply reweight the energy of the true triplets in such a way that the higher-confidence triplets have lower energy than the lower-confidence triplets. As f_{ijp} lies between 0 and 1, we can divide the energy of the true triplets to achieve this reweighting. Thus, the modified loss function becomes

where $[x]_+$ denotes the positive part of x and $y>0$ is a margin hyperparameter and $S'_{(h,l,t)} = \{(h', l, t)|h' \in E\} \cup \{(h, l, t')|t' \in E\}$ is the set of corrupted triplets as described above.

Drug-ADE Prediction

Our first set of experiments followed the procedure described in the study by Mower et al [15] for ADE prediction. Figure 6 shows a schematic of the experiment setting, with details described in the following.

Figure 6. Experiment setting for drug–adverse drug reaction prediction. ADE: adverse drug reaction; CV: cross-validation; ESP: Embedding of Semantic Predications; EU-ADR: Exploring and Understanding Adverse Drug Reactions; KNN: k-nearest neighbors; LR: logistic regression; OMOP: Observational Medical Outcomes Partnership; RF: random forest; TransE: Translating Embeddings for Modeling Multi-relational Data.



Data

We used 2 curated reference data sets that contain drug, disease pairs: Observational Medical Outcomes Partnership (OMOP) [32] and Exploring and Understanding Adverse Drug Reactions (EU-ADR) [8]. OMOP contains 4 ADEs: myocardial infarction, gastrointestinal bleeding, liver injury, and kidney injury for 180 drugs. The drugs for which embeddings from the SemMedDB could not be obtained were removed: 5 in OMOP (corresponding to the drugs darunavir and sitagliptin) and 1 in EU-ADR (for the drug nimesulide). Statistics of both the data sets used in our experiments are presented in Table 1.

ESP embeddings have been empirically evaluated for ADE prediction on these data sets. For each drug and ADE pair, a composite feature vector was obtained by binding the corresponding ESP embeddings. The use of these feature vectors in a logistic regression (LR) classifier was found to outperform previous literature-based methods [15].

Table 1. Exploring and Understanding Adverse Drug Reactions (EU-ADR) and Observational Medical Outcomes Partnership (OMOP) data set statistics.

Data set	Drugs	Diseases	ADE[a] pairs	Non-ADE pairs
EU-ADR	65	10	43	50
OMOP	180	4	164	230

[a]ADE: adverse drug event.

Unsupervised Representation Learning

To compare the performance of the representation learning methods, we generated embeddings of all clinical concepts (nodes) in the SemMedDB using DeepWalk, TransE, Weighted DeepWalk, Weighted TransE, and ESP. We used the available implementation of DeepWalk [33] and TransE [34]. We experimented with multiple hyperparameter sets and selected the ones that yielded the best loss value during representation learning. In Weighted TransE and TransE, we set the α value to .001, batch size to 256 triplets, epochs to 100, and the number of corrupted triplets for each positive triplet to 1. The embedding dimension was 100 in both these models. For the DeepWalk and Weighted DeepWalk models, we set the walk length to 500, the number of walks to 20, window size to 4, α value to .025, and an embedding size of 256. For ESP, we used the embeddings provided by the authors [35], which had a dimension of 8000.

Classification Task and Algorithms

We used supervised binary classification to classify relationships consisting of (drug, disease) pairs. If a drug could cause a particular disease as a side effect, the label assigned to the pair was positive (+1); otherwise, the label assigned was negative (–1). In ESP, the XOR operator is used on each (drug, disease) pair's vector representations (embeddings) to form a single 8000-dimensional input feature vector, which has also been provided by the authors [35]. For all the other embedding algorithms, for each (drug-disease) pair, we concatenated the

vector representations (embeddings) learned from the SemMedDB KG of the drug and disease to form the input feature vectors. The performance of each representation learning technique was evaluated by their classification performance in this task.

To evaluate the downstream effect of the learned representations on classification performance, we used 3 different classifiers. LR is a commonly used linear model. For nonlinear classifiers, we used two different techniques: k-nearest neighbors (KNN) and random forest (RF). KNN classifies a test feature vector based on its distance from the k-nearest training data vectors. RF is an ensemble-based technique that uses multiple decision trees to make predictions. All the experiments in this study were conducted using the scikit-learn library (version 0.24.2) [36]. L1 regularization was used for the LR, information gain (entropy) criterion was used for the RF, and k=5 neighbors was used for the KNN classifier. All other parameters were retained at their default values.

Evaluation

Following the study by Mower et al [15], we evaluated the classifiers using leave-one-out (LOO) cross-validation and stratified 5-fold (S5F) cross-validation on the following data sets:

1. EU-ADR
2. OMOP
3. Combined EU-ADR+OMOP

In LOO cross-validation, the number of folds is equal to the number of instances in the data; in each fold, there is a single test instance, and the remaining instances are used as training data for the classifier. S5F cross-validation is an extension of regular 5-fold cross-validation, where the folds are made by preserving the percentage of samples for each class. In addition, we used two other settings (for a total of 5 settings) to evaluate the generalization performance:

- The classifiers were trained on EU-ADR data and tested on OMOP data.
- The classifiers were trained on OMOP data and tested on EU-ADR data.

Standard metrics to evaluate binary classification were used: the F_1-score and the AUC value. Averages (over all the folds in case of S5F) are reported along with the SD. In the last two settings, we evaluated the trained model on the same 5 folds used in settings 1 and 2, respectively, to have performance values that could be compared.

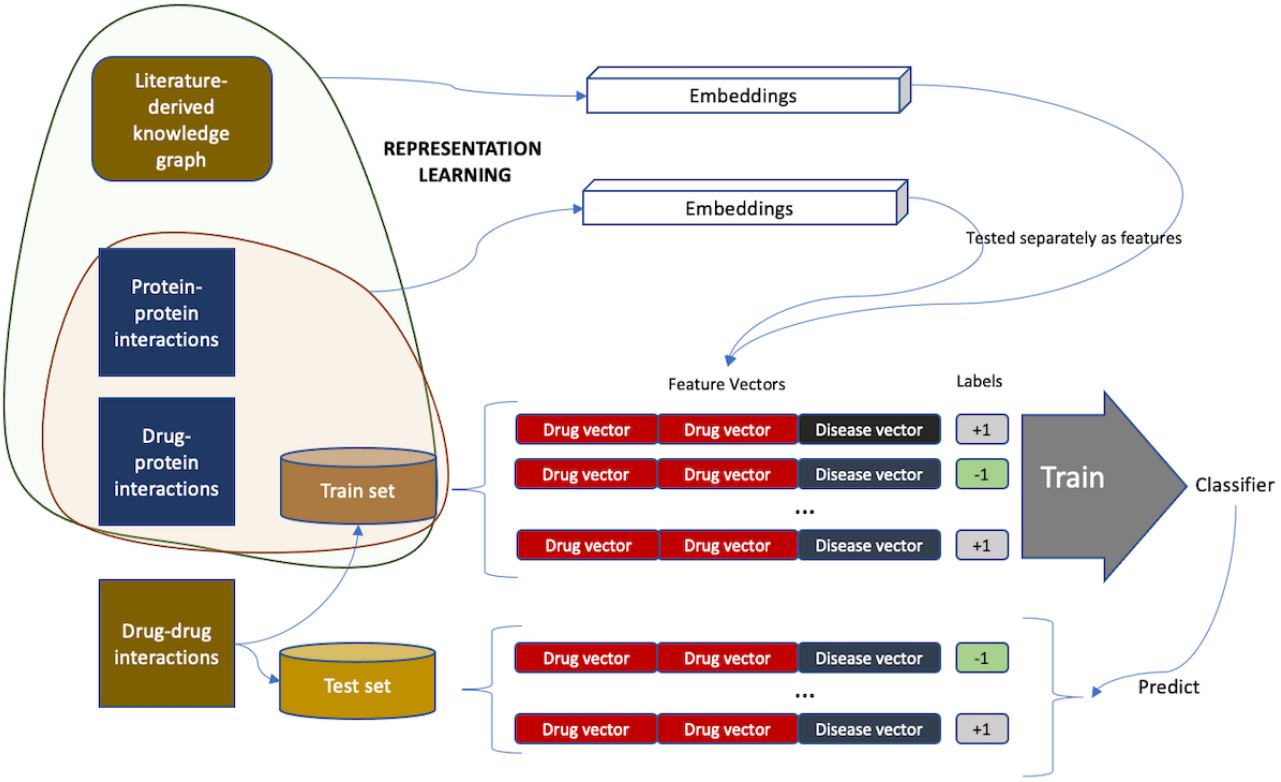
Visualization of Embeddings

To visually inspect the embeddings, we used the dimensionality reduction technique: t-distributed stochastic neighbor embedding (t-SNE) [37]. Embeddings from all 5 methods—for all 487 (drug, disease) pairs in both the OMOP and EU-ADR reference data sets—were plotted in 2 dimensions. The implementation used was from the sklearn library [36]. t-SNE was run with a learning rate of 600. To select the perplexity parameter, we empirically evaluated randomly chosen values and selected those that yielded the best cluster visualization. Perplexity was set to 30 to visualize ESP embeddings and to 5 for all other embeddings. In our plots, we represented the false drug-ADE pairs with an *o* sign and the true drug-ADE pairs with an *x* sign. Each disease was represented by a different color, and there were 10 diseases in total in the data sets.

Polypharmacy Prediction

In this experiment, we evaluated the efficacy of our representation learning methods for polypharmacy prediction. Figure 7 shows a schematic of the experiment setting with details described in the following.

Figure 7. Experiment setting for polypharmacy prediction.



Data

We used the benchmark data set for polypharmacy prediction from the study by Zitnik et al [38], which was also used in the study by Burkhardt et al [39]. We used the same data set, including the exact train-test splits, and called it the polypharmacy data set.

The data consist of 3 types of interactions: drug-drug, drug-protein, and protein-protein interactions. Drug-drug interactions contain triplets of the form drug A-SE-drug B where consuming drug A and drug B together would cause the side effect (SE) mentioned in the triplet, for example, aspirin-Kidney-Failure-warfarin. These data were curated from two databases: Side Effect Resource (SIDER) [40] and Twosides [6]. We used the preprocessed data of the study by Burkhardt et al [39], downloaded from the Zenodo website [41], where triplets for side effects that occurred in fewer than 500 drug interactions are not used. There are 963 side effects in total. Protein-protein interactions, which were curated from multiple databases, indicate physical interactions that have been experimentally found in humans. Drug-protein interactions contain experimentally verified small chemicals (drugs) that target specific proteins. More details can be found in the study by Zitnik et al [38].

The drug-drug interaction data were divided into 80% training, 10% validation, and 10% test sets. Furthermore, along with valid interactions, that is, where the 2 drugs cause the reported side effect, the study by Zitnik et al [38] also provides an equal number of invalid interactions by using randomly selected drugs and side effects that do not occur in the valid interactions. There are 22,89,960 protein-protein interactions and 29,756 drug-protein interactions that were used only during training. Textbox 1 shows the number of interactions provided in the benchmark data set.

Textbox 1. Polypharmacy data set statistics.

Number of interactions provided in the benchmark data set

- Train
  - Drug-drug interactions: 73,23,790
  - Protein-protein interactions (only used during training): 22,89,960
  - Drug-protein interactions (only used during training): 29,756
- Test
  - Valid drug-drug interactions (label 1): 4,57,196
  - Invalid drug-drug interactions (label 0): 4,57,196

The previous best results published on this data set, to our knowledge, are those of ESP in the study by Burkhardt et al [39]. The authors' approach first learns ESP embeddings from drug-drug, drug-protein, and protein-protein interactions of the polypharmacy training data set. Given a test triple—drug 1, drug 2, and side effect—they bind the embeddings of the drugs, and if the similarity of the obtained composite vector to the side effect embedding is more than a fixed threshold, they predict the triple to be valid. This simple approach was found to outperform Decagon, a more complex graph neural network–based approach [38].

Unsupervised Representation Learning

We generated embeddings in two different ways. First, we used the training data provided to learn embeddings using TransE, Weighted TransE, DeepWalk, and Weighted DeepWalk. We used the available implementations of DeepWalk [33] and TransE [34]. We experimented with multiple hyperparameter sets and selected the ones that yielded the best loss value during representation learning. Both DeepWalk and Weighted DeepWalk embeddings were generated by setting the number of walks to 25, walk length to 500, window size to 10, embedding size to 256, and α value to .025. TransE and Weighted TransE embeddings were generated by setting batch size to 512, number of corrupted triplets for each positive triplet to 1, epochs to 100, α value to .001, and embedding size to 100. In the weighted versions, the occurrence score for a triplet drug A-SE-drug B is the number of triplets containing the same drug A (subject) and drug B (object) with no restrictions on the side effect (ie, they may have different side effects), and subject, object scores were set to 1.

Second, to evaluate the utility of the SemMedDB as another auxiliary data source (in addition to protein-protein interaction and drug-protein interaction networks), we augmented the training data with 93,974,376 triplets from the SemMedDB. The subject, object scores and co-occurrence scores of the SemMedDB were reused as such for the SemMedDB; for the polypharmacy data set, the subject, object scores and co-occurrence scores were set to the highest values found in the SemMedDB (which were 1000, 1000, and 33,478, respectively). As this was a much larger graph, different hyperparameters were used to obtain embeddings. DeepWalk embeddings were generated by setting the number of walks to 375, walk length to 500, window size to 10, embedding size to 256, and α value to .025. TransE and Weighted TransE embeddings were generated by setting batch size to 512, number of corrupted triplets for each positive triplet to 2, epochs to 1500, α value to .001, and embedding size to 100.

Classification Task and Settings

The binary classification task was to distinguish the valid and invalid drug-drug interactions (in the test set provided). After the embeddings were learned (separately in the two settings), we trained an RF classifier from sklearn—with the number of decision trees set to 100 and maximum depth of each tree set to 20 (all other settings were unchanged from the default)—for this task. For each drug-SE-drug interaction, the embeddings of both the drugs and the side effect were concatenated and used as a feature vector in the classifier. This concatenation yielded a 300-dimensional feature vector for each triplet in the case of TransE and Weighted TransE and a 768-dimensional feature vector in the case of DeepWalk and Weighted DeepWalk.

As there were no invalid interactions in the training set, we randomly generated 73,23,790 pairs of drug-drug interactions such that they did not occur in either the training set or test set provided in the benchmark data.

Evaluation Metrics

The evaluation metrics used were the same as the ones used in the studies by Zitnik et al [38] and Burkhardt et al [39]: AUC, area under the precision-recall curve (AUPRC), and average precision at 50 (AP@50) for each of the 963 side effects, which were then averaged. We compared our results with the published results reported on the same data set in the study by Burkhardt et al [39], which used ESP-based embeddings, and in the study by Zitnik et al [38], which used Decagon, a graph convolutional network developed for this task.

Results

Drug-ADE Prediction

Tables 2-4 show the F_1-scores on the LOO and S5F cross-validation configurations obtained by the algorithms for the data sets OMOP, EU-ADR, and the combined OMOP+EU-ADR data set, respectively. In most cases, we observed that ESP outperformed TransE and DeepWalk. However, both the weighted versions—Weighted TransE and Weighted DeepWalk—outperformed ESP in most cases. Among the 3 classifiers, for the same embeddings, RF outperformed LR and KNN in most cases. Overall, Weighted TransE with RF had the best performance in most cases, with improvements of up to 5.75% over ESP.

Table 2. F_1-scores from leave-one-out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations on the Observational Medical Outcomes Partnership data set.

Model	ESP[a]	TransE[b]	DeepWalk	Weighted DeepWalk	Weighted TransE	Increase (%)[c]
LR[d] S5F, mean (SD)	0.895 (0.02)	0.861 (0.0185)	0.813 (0.024)	0.899 (0.0186)	0.915[e] (0.0178)	2.23
LR LOO-CV	0.901	0.889	0.828	0.912	0.923[e]	2.44
KNN[f] S5F, mean (SD)	0.816 (0.016)	0.793 (0.0163)	0.784 (0.0173)	0.814 (0.0167)	0.838[e] (0.0155)	2.69
KNN LOO-CV	0.837	0.804	0.796	0.837	0.859[e]	2.63
RF[g] S5F, mean (SD)	0.906 (0.008)	0.865 (0.0078)	0.82 (0.0091)	0.91 (0.0077)	0.923[e] (0.0069)	1.87
RF LOO-CV	0.921	0.877	0.834	0.931	0.936[e]	1.69

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications (in each row).

[d]LR: logistic regression.

[e]Best result in each row.

[f]KNN: k-nearest neighbors.

[g]RF: random forest.

Table 3. F_1-scores from leave-one-out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations on the Exploring and Understanding Adverse Drug Reactions data set.

Model	ESP[a]	TransE[b]	DeepWalk	Weighted DeepWalk	Weighted TransE	Increase (%)[c]
LR[d] S5F, mean (SD)	0.834 (0.066)	0.823 (0.073)	0.769 (0.089)	0.832 (0.068)	0.857[e] (0.0635)	2.76
LR LOO-CV	0.841	0.827	0.783	0.843	0.864[e]	2.73
KNN[f] S5F, mean (SD)	0.621 (0.085)	0.55 (0.096)	0.646[e] (0.076)	0.639 (0.079)	0.643 (0.0732)	4.02
KNN LOO-CV	0.641	0.651	0.659	0.665[e]	0.663	3.74
RF[g] S5F, mean (SD)	0.835 (0.017)	0.824 (0.022)	0.77 (0.035)	0.856[e] (0.0157)	0.855 (0.0162)	2.51
RF LOO-CV	0.85	0.826	0.785	0.874	0.877[e]	3.17

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

[d]LR: logistic regression.

[e]Best result in each row.

[f]KNN: k-nearest neighbors.

[g]RF: random forest.

Table 4. F_1-scores from leave-one-out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations on combined Observational Medical Outcomes Partnership+Exploring and Understanding Adverse Drug Reactions data set.

Model	ESP[a]	TransE[b]	DeepWalk	Weighted DeepWalk	Weighted TransE	Increase (%)[c]
LR[d] S5F, mean (SD)	0.886 (0.021)	0.855 (0.033)	0.843 (0.039)	0.897 (0.025)	0.904[e] (0.0203)	2.03
LR LOO-CV	0.911	0.871	0.856	0.928	0.934[e]	2.52
KNN[f] S5F, mean (SD)	0.817 (0.035)	0.768 (0.043)	0.755 (0.049)	0.831 (0.047)	0.836[e] (0.032)	2.32
KNN LOO-CV	0.829	0.79	0.776	0.842	0.848[e]	2.29
RF[g] S5F, mean (SD)	0.86 (0.021)	0.86 (0.024)	0.845 (0.025)	0.892 (0.021)	0.898[e] (0.0208)	4.42
RF LOO-CV	0.87	0.868	0.862	0.898	0.92[e]	5.75

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

[d]LR: logistic regression.

[e]Best result in each row.

[f]KNN: k-nearest neighbors.

[g]RF: random forest.

Tables 5-7 show the AUC values for the LOO and S5F cross-validation configurations obtained by the algorithms for the data sets OMOP, EU-ADR, and the combined OMOP+EU-ADR data set, respectively. We observed the same trend that was observed with the F_1-scores. After accounting for the weights, the results for both DeepWalk and TransE improved over those of ESP. On the OMOP data set, LR with Weighted TransE obtained the highest improvement of 3.85%. On the EU-ADR data set, KNN performed the best and with both the weighted versions improved over ESP by approximately 8%. On the combined data set, KNN with Weighted TransE had the highest improvement over ESP.

Table 5. Area under the receiver operating characteristic curve values from leave-one-out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations on the Observational Medical Outcomes Partnership data set.

Model	ESP[a]	TransE[b]	DeepWalk	Weighted DeepWalk	Weighted TransE	Increase (%)[c]
LR[d] S5F, mean (SD)	0.935 (0.024)	0.935 (0.022)	0.892 (0.026)	0.932 (0.0287)	0.971[e] (0.023)	3.85
LR LOO-CV	0.94	0.938	0.901	0.963	0.965[e]	2.66
KNN[f] S5F, mean (SD)	0.883 (0.023)	0.858 (0.024)	0.841 (0.0267)	0.891 (0.021)	0.911[e] (0.027)	3.17
KNN LOO-CV	0.902	0.875	0.857	0.894	0.924[e]	2.44
RF[g] S5F, mean (SD)	0.945 (0.008)	0.931 (0.0091)	0.888 (0.0078)	0.958 (0.0077)	0.971[e] (0.0069)	2.75
RF LOO-CV	0.961	0.943	0.881	0.971	0.972[e]	1.14

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

[d]LR: logistic regression.

[e]Best result in each row.

[f]KNN: k-nearest neighbors.

[g]RF: random forest.

Table 6. Area under the receiver operating characteristic curve values from leave-one-out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations on the Exploring and Understanding Adverse Drug Reactions data set.

Model	ESP[a]	TransE[b]	DeepWalk	Weighted DeepWalk	Weighted TransE	Increase (%)[c]
LR[d] S5F, mean (SD)	0.885 (0.085)	0.897 (0.076)	0.825 (0.0732)	0.901 (0.096)	0.929[e] (0.0743)	4.97
LR LOO-CV	0.903	0.899	0.843	0.902	0.919[e]	1.77
KNN[f] S5F, mean (SD)	0.693 (0.076)	0.634 (0.072)	0.753 (0.074)	0.702 (0.083)	0.712[e] (0.087)	8.66
KNN LOO-CV	0.721	0.734	0.778	0.784[e]	0.752	8.74
RF[g] S5F, mean (SD)	0.894 (0.0164)	0.897 (0.089)	0.826 (0.068)	0.924[e] (0.0635)	0.924[e] (0.066)	3.36
RF LOO-CV	0.922	0.901	0.847	0.927	0.933[e]	1.19

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

[d]LR: logistic regression.

[e]Best result in each row.

[f]KNN: k-nearest neighbors.

[g]RF: random forest.

Table 7. Area under the receiver operating characteristic curve values from leave-one-out (LOO) and stratified 5-fold (S5F) cross-validation (CV) configurations on the combined Observational Medical Outcomes Partnership+Exploring and Understanding Adverse Drug Reactions data set.

Model	ESP[a]	TransE[b]	DeepWalk	Weighted DeepWalk	Weighted TransE	Increase (%)[c]
LR[d] S5F, mean (SD)	0.932 (0.027)	0.925 (0.021)	0.901 (0.026)	0.93 (0.024)	0.945[e] (0.023)	1.39
LR LOO-CV	0.952	0.935	0.923	0.969	0.972[e]	2.1
KNN[f] S5F, mean (SD)	0.885 (0.033)	0.823 (0.025)	0.806 (0.0203)	0.901 (0.0227)	0.906[e] (0.039)	2.37
KNN LOO-CV	0.898	0.855	0.837	0.9	0.923[e]	2.78
RF[g] S5F, mean (SD)	0.92 (0.038)	0.93 (0.042)	0.9 (0.025)	0.937 (0.0247)	0.94[e] (0.026)	2.17
RF LOO-CV	0.951[e]	0.92	0.929	0.935	0.94	–1.1

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

[d]LR: logistic regression.

[e]Best result in each row.

[f]KNN: k-nearest neighbors.

[g]RF: random forest.

Table 8 shows the performance of the methods when the embeddings and classifiers were trained on OMOP data and prediction was carried out on EU-ADR data. Table 9 shows the results of the opposite case: training on EU-ADR data and prediction on OMOP data. Compared with the values in Tables 2-7, we observed lower F_1-scores and AUC values in general. As the test set was from a different source, it was harder for the trained models to generalize. The performance trend across the algorithms remained the same: DeepWalk and TransE did not outperform ESP, but our weighted versions outperformed ESP. Weighted TransE had the best performance in most cases, with improvements of up to 8.4% in AUC value and 3.7% in the F_1-score with LR (as seen in Table 9).

Table 8. Area under the receiver operating characteristic curve (AUC) values and F_1-scores: training on Observational Medical Outcomes Partnership data and prediction on Exploring and Understanding Adverse Drug Reactions data.

Model and metric	ESP[a], mean (SD)	TransE[b], mean (SD)	DeepWalk, mean (SD)	Weighted DeepWalk, mean (SD)	Weighted TransE, mean (SD)	Increase (%)[c]
Logistic regression						
F_1	0.715 (0.023)	0.711 (0.026)	0.703 (0.031)	0.734 (0.037)	0.737[d] (0.027)	3.07
AUC	0.798 (0.017)	0.788 (0.023)	0.769 (0.0164)	0.803[d] (0.018)	0.802 (0.019)	0.63
KNN[e]						
F_1	0.712 (0.028)	0.702 (0.038)	0.698 (0.042)	0.729 (0.029)	0.734[d] (0.024)	3.09
AUC	0.785 (0.022)	0.767 (0.021)	0.764 (0.023)	0.801 (0.019)	0.804[d] (0.027)	2.42
Random forest						
F_1	0.724 (0.036)	0.714 (0.039)	0.710 (0.041)	0.745 (0.022)	0.748[d] (0.021)	3.31
AUC	0.815 (0.007)	0.800 (0.008)	0.783 (0.005)	0.818 (0.006)	0.825[d] (0.007)	1.23

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

[d]Best result in each row.

[e]KNN: k-nearest neighbors.

Table 9. Area under the receiver operating characteristic curve (AUC) scores and F_1-scores: training on Exploring and Understanding Adverse Drug Reactions data and prediction on Observational Medical Outcomes Partnership data.

Model and metric	ESP[a], mean (SD)	TransE[b], mean (SD)	DeepWalk, mean (SD)	Weighted DeepWalk, mean (SD)	Weighted TransE, mean (SD)	Increase (%)[c]
Logistic regression						
F-1	0.612 (0.018)	0.604 (0.028)	0.597 (0.034)	0.632 (0.019)	0.635[d] (0.021)	3.76
AUC	0.680 (0.028)	0.678 (0.019)	0.666 (0.022)	0.739[d] (0.025)	0.737 (0.021)	8.67
KNN[e]						
F-1	0.609 (0.028)	0.598 (0.036)	0.589 (0.033)	0.628 (0.25)	0.633[d] (0.0223)	3.94
AUC	0.684 (0.004)	0.665 (0.008)	0.648 (0.003)	0.731 (0.006)	0.734[d] (0.007)	7.31
Random forest						
F-1	0.630 (0.032)	0.617 (0.035)	0.601 (0.043)	0.641 (0.029)	0.651[d] (0.0286)	3.33
AUC	0.717 (0.022)	0.685 (0.017)	0.675 (0.019)	0.750 (0.023)	0.763[d] (0.027)	6.41

[a]ESP: Embedding of Semantic Predications.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]Improvement in performance of the best method over Embedding of Semantic Predications.

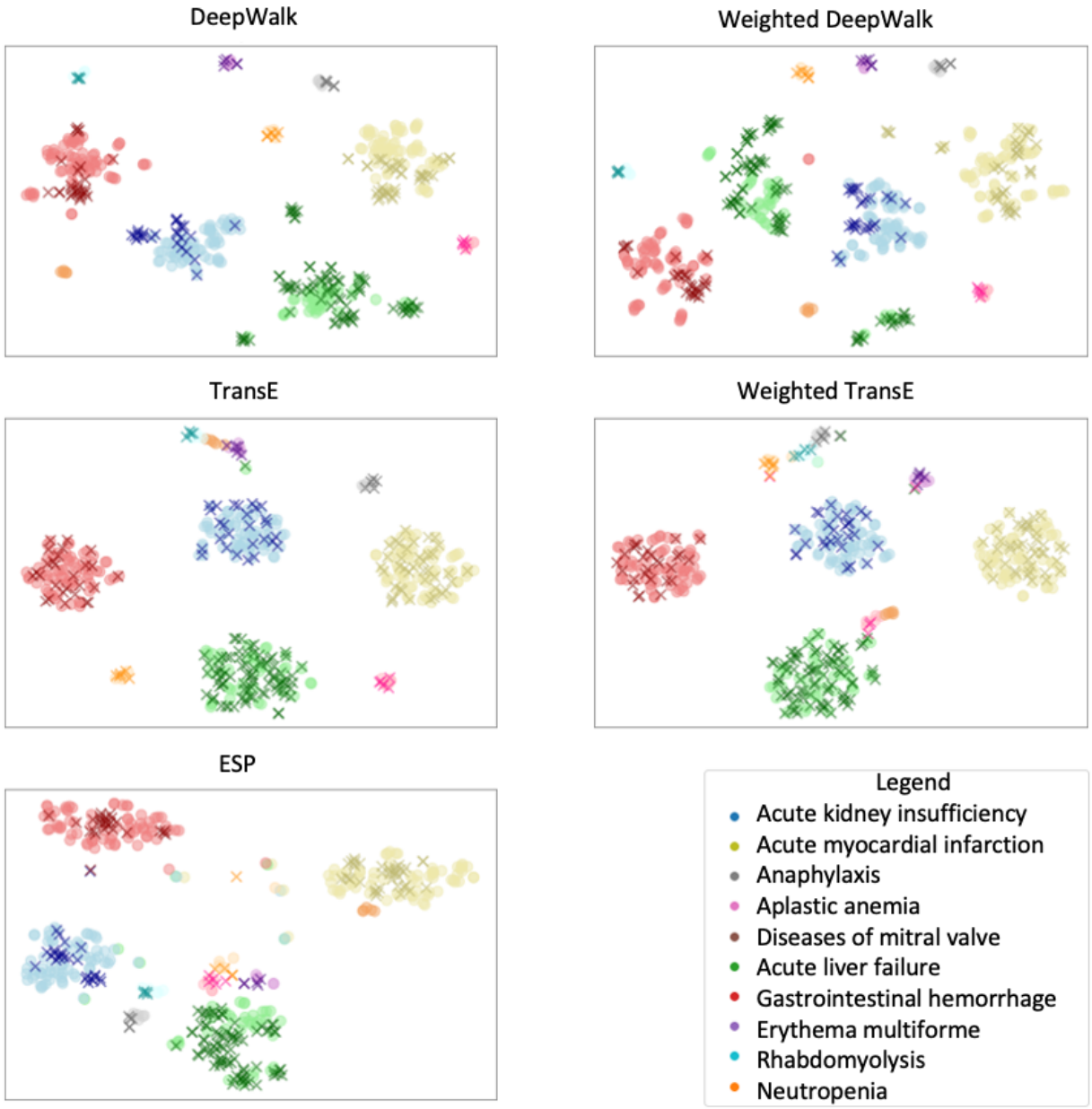
[d]Best result in each row.

[e]KNN: k-nearest neighbors.

Visualization

Figure 8 shows the t-SNE visualizations of the embeddings obtained from each of the methods. The cluster structure with respect to the diseases is clear in all the embeddings. The intercluster separation seems to be the best for TransE and its weighted version, where the clusters (after the t-SNE dimensionality reduction) are also more compactly distributed. Within each cluster, the positive and negative instances do not appear to be well separated, although there is some localization seen in the ESP, DeepWalk, and Weighted DeepWalk clusters. The results in the previous sections show that, among the classifiers we tested, RF and KNN had the best performance. Both indicated that the boundary between positive and negative instances was nonlinear.

Figure 8. t-distributed stochastic neighbor embedding plots of disease-drug embeddings of the combined Exploring and Understanding Adverse Drug Reactions+Observational Medical Outcomes Partnership data set. Color indicates disease, and the markers x and o indicate presence and absence, respectively, of a side effect. ESP: Embedding of Semantic Predications; TransE: Translating Embeddings for Modeling Multi-relational Data.



Polypharmacy Prediction

We first used TransE, DeepWalk, and their weighted versions to obtain embeddings from the data (without the use of the SemMedDB). As shown in Table 10, the performance of all 4 methods was superior to that of ESP in terms of the mean AUC value and mean AUPRC value. DeepWalk outperformed TransE. The weighted versions were not superior presumably because the underlying graph was not noisy and the weights based on co-occurrences alone in the drug-drug interaction graph did not affect the representation learning.

When the SemMedDB was added as an additional data source from which to learn embeddings, the results improved, as shown in Table 10. The maximum increase was seen in the AP@50 metric for both TransE and DeepWalk. This shows that the addition of the SemMedDB during representation learning improves the precision of the classifier learned. The weighted versions were not significantly better than the corresponding unweighted versions because the Decagon graphs had a significantly higher number of triplets than the SemMedDB, which dominated the data.

Overall, the representations learned with Weighted DeepWalk on the SemMedDB and polypharmacy graphs (drug-drug, drug-protein, and protein-protein interactions), when used in the RF classifier obtained the best results advancing the state of the art by 3.5% in the mean AUC value, 4.3% in the mean AUPRC value, and 5% in the mean AP@50 value.

Table 10. Mean (SD) area under the receiver operating characteristic curve (AUC), area under the precision-recall curve (AUPRC), and average precision at 50 (AP@50) values (averaged over 963 side effects) on the polypharmacy data set.

Metric	Polypharmacy graphs, mean (SD)				SemMedDB[a]+polypharmacy graphs, mean (SD)				Published results	
	TransE[b]	Weighted TransE	DeepWalk	Weighted DeepWalk	TransE	Weighted TransE	DeepWalk	Weighted DeepWalk	ESP[c]	Decagon
Mean AUC	0.921 (0.021)	0.921 (0.021)	0.932 (0.019)	0.932 (0.019)	0.926 (0.024)	0.924 (0.024)	0.935[d] (0.020)	0.935[d] (0.020)	0.903 (0.023)	0.872
Mean AUPRC	0.877 (0.037)	0.875 (0.038)	0.91 (0.030)	0.911 (0.030)	0.906 (0.033)	0.904 (0.033)	0.912 (0.031)	0.913[d] (0.031)	0.875 (0.034)	0.832
Mean AP@50	0.73 (0.165)	0.725 (0.167)	0.896 (0.082)	0.896 (0.081)	0.915 (0.061)	0.916 (0.061)	0.906 (0.066)	0.909[d] (0.064)	0.865 (0.073)	0.803

[a]SemMedDB: Semantic MEDLINE Database.

[b]TransE: Translating Embeddings for Modeling Multi-relational Data.

[c]ESP: Embedding of Semantic Predications.

[d]Best result in each row.

Discussion

Principal Findings

Unsupervised representation learning enables us to find useful features from data without requiring task-specific labels, which can subsequently be used in multiple applications. This is particularly useful when labeled data for a specific task are scarce, such as in ADE prediction, and when the data are complex, which is the case for KGs. Biomedical KGs such as the SemMedDB are inferred from the literature through NLP. This inference process introduces noise in the form of erroneous or incomplete edges and nodes in the KG. We developed new techniques to model underlying noise during representation learning from literature-derived biomedical KGs. During NLP inference, confidence scores were assigned to the inferred clinical concepts (vertices) and relations (edges). Our method effectively used these confidence scores during representation learning to model the inaccuracies in the graphs due to NLP inference.

We illustrated the use of our technique on two well-known representation learning methods: DeepWalk and TransE. We showed how confidence scores can easily be incorporated in both these methods to develop their *weighted* versions: Weighted DeepWalk and Weighted TransE. We compared the performance of these methods with ESP, which is, to our knowledge, the best-known representation learning method designed for the SemMedDB, a literature-derived KG. All the experiments were performed on benchmark data sets for ADE prediction.

In one set of experiments, the drug and disease embeddings learned from various representation learning methods were used to train classifiers and predict on held-out test sets in various cross-validation configurations. In another set of experiments, the side effects of drug-drug interactions were predicted using other drug-drug interactions as well as auxiliary data on drug-protein and protein-protein interactions. In the latter case, representations were learned both with and without the use of KGs. In both sets of experiments, our weighted versions learned representations that yielded more accurate predictive models than ESP as well as the unweighted versions of DeepWalk and TransE. Visual inspection of the learned embeddings shows a clear cluster structure in compressed 2-dimensional view, indicating that the disease and drug embeddings have been learned well from the KG.

In the second set of experiments, the use of biomedical KG as an auxiliary data source was found to considerably improve the precision. When the KG was not used as an auxiliary source, our weighted versions did not outperform the unweighted versions of DeepWalk and TransE for representation learning from drug-drug, drug-protein, and protein-protein interaction graphs. These graphs are not literature-derived, and the weights were based on co-occurrence scores in lieu of confidence scores. This shows that when the underlying graphs are not noisy, the weights may not add much value, although the performance does not deteriorate.

Our weighted versions of DeepWalk and TransE are, by design, biased toward triplets that have high co-occurrence scores in literature-derived KGs. This may not favor *some* relations that have low co-occurrence scores. The low score may be due to the triplet being a recently discovered relation or because it may be mentioned infrequently in the literature. However, the aim of graph representation learning methods is to use the entire KG, including indirectly related concepts, to learn the representation of a clinical concept. Therefore, if there are other (older or more frequent) relations that strongly indicate the possibility of the relation with low co-occurrence, then this signal is captured during representation learning. It is exactly this ability of graph representation learning that makes it useful in link prediction for knowledge discovery [20].

To evaluate this in our specific context, we checked the predictive accuracy of our weighted approach using the EU-ADR data set for those true drug-ADE pairs that have relations with low co-occurrence scores in the SemMedDB. Table 11 lists the predicates and their co-occurrence scores for 3 drug (subject) and disease (object) pairs. Note that these co-occurrence scores are much smaller than the maximum value of 33,478.

Table 11. Drug, adverse drug event (ADE) pairs from the Exploring and Understanding Adverse Drug Reactions data set with low co-occurrence scores in the Semantic MEDLINE Database.

Drug	ADE	Predicates (co-occurrence score)
Diclofenac	Anaphylaxis	Causes (6), Predisposes (1)
Aspirin	Anaphylaxis	Disrupts (2), Augments (2), Affects (2), Causes (7), Treats (1)
Acetaminophen	Anaphylaxis	Causes (7), Treats (3), Affects (1)

We checked the predictions for each of the aforementioned 3 pairs using classifiers trained on the SemMedDB representations generated using Weighted DeepWalk. In all, 3 classifiers—RF, KNN, and LR—were trained on EU-ADR data after excluding the pair being tested. The features were obtained by concatenating the corresponding drug and disease representations, as was done for the experiments on drug-ADE prediction. All 3 classifiers correctly identified the 3 pairs as true positives. This strongly suggests that the representations could learn the indirect relations from the KG despite being biased through our weighted approach toward relations with high co-occurrence scores.

To summarize, all our experimental results clearly highlight the importance of modeling inaccuracies in the inferred KGs for representation learning.

Limitations

This study has the following limitations. Our weighting technique relies on the confidence scores provided and thus, in turn, depends on the accuracy of these scores. Errors in these scores may be detrimental to representation learning, and their effects need to be evaluated further. Model designers should be aware of this limitation when such weighting schemes are used in other KGs.

We evaluated the use of our weighting scheme on just two representation learning methods: DeepWalk and TransE. Many other methods exist, especially for heterogeneous networks; a recent survey can be found in the study by Yang et al [27]. Despite the simplicity of these approaches, we obtained very good results, outperforming the state of the art for ADE prediction. We believe that the underlying idea of our weighting scheme can be applied to many other representation learning methods, which can be investigated in the future.

In our experiments, both OMOP and EU-ADR data sets were not large. Although in our experiments, we rigorously tested many cross-validation configurations, accuracy values can differ in other data sets. We also note that the reported performance was dependent on the KG used to learn representations from, and the results may vary with other KGs. This is less of a concern for the second polypharmacy data set, which was much larger. The relative performances of the methods showed a consistent trend across both data sets. Comparisons with more diverse data sets will further our understanding of the strengths and limitations of these methods.

Future Work

This work can be extended in many ways. Alternative approaches to designing the scoring function and weighting scheme used in our weighting function can be investigated. In large data sets, it may be possible to learn the weights automatically from the data by suitably modifying the models. The weighting scheme can be extended to incorporate additional information in literature-derived KGs. To leverage the underlying biomedical literature used, techniques to obtain causal assessment can also be explored.

Additional experiments can be designed to compare our approach with a fully supervised approach where both the embeddings and the classifier are learned jointly. Future work can also evaluate the effects of KG characteristics on the performance by experimenting with other KGs. Finally, the utility of our representations in other tasks such as diagnosis prediction or finding new clinical associations can also be evaluated.

Conclusions

Literature-derived KGs are an important resource for analyzing the wealth of knowledge stored in the growing biomedical literature. These KGs are inferred through NLP techniques, and their limitations may result in incomplete or erroneous nodes and edges. KG embeddings provide a scalable and automatic way of obtaining features from KGs that can be valuable in multiple biomedical prediction tasks. Our work demonstrates the need for modeling noise in the underlying KG and makes an important step toward improved representation learning from literature-derived KGs and thus toward effectively using literature-derived KGs for predictive models.

Our experiments show that such *noise-aware* representations in turn lead to classifiers for ADE prediction that are more accurate than representations learned from the best previous methods. The new models in this work can be used by pharmacovigilance teams to detect previously unknown ADEs for further evaluation. Software implementation of our new methods and all experiments are publicly available at our website [42].

Acknowledgments

This work is supported by the Singapore Ministry of Education Academic Research Fund [R-253-000-138-133], primary investigator: VR.

Authors' Contributions

SD, ALJH, RM, and VR designed the algorithms. ALJH implemented Weighted DeepWalk. SD implemented Weighted TransE and the scripts to run the experiments on ADE prediction. AJ implemented the scripts to run experiments on polypharmacy prediction. SD and AJ conducted all the computational experiments. VR, SD, AJ, and RM wrote the manuscript. VR conceived and supervised the project.

Conflicts of Interest

None declared.

References

1. Ventola CL. Big data and pharmacovigilance: data mining for adverse drug events and interactions. P T 2018 Jun;43(6):340-351 [FREE Full text] [Medline: 29896033]
2. Watanabe JH, McInnis T, Hirsch JD. Cost of prescription drug-related morbidity and mortality. Ann Pharmacother 2018 Sep;52(9):829-837. [doi: 10.1177/1060028018765159] [Medline: 29577766]
3. Downing NS, Shah ND, Aminawung JA, Pease AM, Zeitoun J, Krumholz HM, et al. Postmarket safety events among novel therapeutics approved by the US food and drug administration between 2001 and 2010. J Am Med Assoc 2017 May 09;317(18):1854-1863 [FREE Full text] [doi: 10.1001/jama.2017.5150] [Medline: 28492899]
4. Mokhtari RB, Homayouni TS, Baluch N, Morgatskaya E, Kumar S, Das B, et al. Combination therapy in combating cancer. Oncotarget 2017 Jun 06;8(23):38022-38043 [FREE Full text] [doi: 10.18632/oncotarget.16723] [Medline: 28410237]
5. Kantor ED, Rehm CD, Haas JS, Chan AT, Giovannucci EL. Trends in prescription drug use among adults in the United States from 1999-2012. J Am Med Assoc 2015 Nov 03;314(17):1818-1831 [FREE Full text] [doi: 10.1001/jama.2015.13766] [Medline: 26529160]
6. Tatonetti NP, Ye PP, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Transl Med 2012 Mar 14;4(125):125ra31 [FREE Full text] [doi: 10.1126/scitranslmed.3003377] [Medline: 22422992]
7. Percha B, Altman RB. Informatics confronts drug-drug interactions. Trends Pharmacol Sci 2013 Mar;34(3):178-184 [FREE Full text] [doi: 10.1016/j.tips.2013.01.006] [Medline: 23414686]
8. Coloma PM, Avillach P, Salvo F, Schuemie MJ, Ferrajolo C, Pariente A, et al. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. Drug Saf 2013 Jan;36(1):13-23. [doi: 10.1007/s40264-012-0002-x] [Medline: 23315292]
9. Du L, Chakraborty A, Chiang C, Cheng L, Quinney S, Wu H, et al. Graphic mining of high-order drug interactions and their directional effects on myopathy using electronic medical records. CPT Pharmacometrics Syst Pharmacol 2015 Aug;4(8):481-488 [FREE Full text] [doi: 10.1002/psp4.59] [Medline: 26380157]
10. Vilar S, Friedman C, Hripcsak G. Detection of drug-drug interactions through data mining studies using clinical sources, scientific literature and social media. Brief Bioinform 2018 Sep 28;19(5):863-877 [FREE Full text] [doi: 10.1093/bib/bbx010] [Medline: 28334070]
11. MEDLINE: Overview. National Library of Medicine. URL: https://www.nlm.nih.gov/medline/medline_overview.html [accessed 2021-09-24]
12. Cohen T, Widdows D, Schvaneveldt RW, Davies P, Rindflesch TC. Discovering discovery patterns with predication-based semantic indexing. J Biomed Inform 2012 Dec;45(6):1049-1065 [FREE Full text] [doi: 10.1016/j.jbi.2012.07.003] [Medline: 22841748]
13. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Saf 2014 Oct;37(10):777-790 [FREE Full text] [doi: 10.1007/s40264-014-0218-z] [Medline: 25151493]
14. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc 2012 Jun;19(e1):28-35 [FREE Full text] [doi: 10.1136/amiajnl-2011-000699] [Medline: 22718037]
15. Mower J, Subramanian D, Cohen T. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. J Am Med Inform Assoc 2018 Oct 01;25(10):1339-1350 [FREE Full text] [doi: 10.1093/jamia/ocy077] [Medline: 30010902]
16. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 2005 Dec;6(4):357-369. [doi: 10.1093/bib/6.4.357] [Medline: 16420734]
17. McIntosh T, Curran JR. Challenges for automatically extracting molecular interactions from full-text articles. BMC Bioinformatics 2009 Sep 24;10:311 [FREE Full text] [doi: 10.1186/1471-2105-10-311] [Medline: 19778419]
18. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 2003 Dec;36(6):462-477 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.003] [Medline: 14759819]
19. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014 Presented at: KDD '14: The 20th

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 24 - 27, 2014; New York New York USA p. 701-710. [doi: 10.1145/2623330.2623732]
20. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modelling multi-relational data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: NIPS'13: 26th International Conference on Neural Information Processing Systems; December 5 - 10, 2013; Lake Tahoe Nevada p. 2787-2795 URL: https://dl.acm.org/doi/10.5555/2999792.2999923
21. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):267-270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
22. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics 2012 Dec 01;28(23):3158-3160 [FREE Full text] [doi: 10.1093/bioinformatics/bts591] [Medline: 23044550]
23. Ernst P, Siu A, Weikum G. KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences. BMC Bioinformatics 2015 May 14;16:157 [FREE Full text] [doi: 10.1186/s12859-015-0549-5] [Medline: 25971816]
24. Gopalakrishnan V, Jha K, Jin W, Zhang A. A survey on literature based discovery approaches in biomedical domain. J Biomed Inform 2019 May;93:103141 [FREE Full text] [doi: 10.1016/j.jbi.2019.103141] [Medline: 30857950]
25. Bakal G, Talari P, Kakani EV, Kavuluru R. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. J Biomed Inform 2018 Jun;82:189-199 [FREE Full text] [doi: 10.1016/j.jbi.2018.05.003] [Medline: 29763706]
26. Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: a survey. Knowl Based Syst 2018 Jul 01;151:78-94. [doi: 10.1016/j.knosys.2018.03.022]
27. Yang C, Xiao Y, Zhang Y, Sun Y, Han J. Heterogeneous network representation learning: a unified framework with survey and benchmark. IEEE Trans Knowl Data Eng 2020 Dec 21:1-1. [doi: 10.1109/tkde.2020.3045924]
28. Wang Q, Mao Z, Wang B, Guo L. Knowledge graph embedding: a survey of approaches and applications. IEEE Trans Knowl Data Eng 2017 Sep 20;29(12):2724-2743. [doi: 10.1109/tkde.2017.2754499]
29. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: NIPS'13: 26th International Conference on Neural Information Processing Systems; December 5 - 10, 2013; Lake Tahoe Nevada p. 3111-3119. [doi: 10.5555/2999792.2999959]
30. Cohen T, Widdows D. Embedding of semantic predications. J Biomed Inform 2017 Apr;68:150-166 [FREE Full text] [doi: 10.1016/j.jbi.2017.03.003] [Medline: 28284761]
31. Dong Y, Chawla N, Swami A. metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 Presented at: KDD '17: The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13 - 17, 2017; Halifax NS Canada p. 135-144. [doi: 10.1145/3097983.3098036]
32. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf 2013 Oct;36 Suppl 1:33-47. [doi: 10.1007/s40264-013-0097-8] [Medline: 24166222]
33. DeepWalk - Deep learning for graphs. Github. URL: https://github.com/phanein/deepwalk [accessed 2021-09-24]
34. An open-source package for Knowledge Embedding (KE). Github. URL: https://github.com/thunlp/OpenKE/tree/OpenKE-Tensorflow1.0 [accessed 2021-09-24]
35. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. Github. URL: https://github.com/jusger/ADEClassifier-RepLearnML [accessed 2021-09-24]
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
37. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605 [FREE Full text]
38. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 2018 Jul 01;34(13):457-466 [FREE Full text] [doi: 10.1093/bioinformatics/bty294] [Medline: 29949996]
39. Burkhardt HA, Subramanian D, Mower J, Cohen T. Predicting adverse drug-drug interactions with neural embedding of semantic predications. AMIA Annu Symp Proc 2020 Mar 04;2019:992-1001 [FREE Full text] [Medline: 32308896]
40. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016 Jan 04;44(D1):1075-1079 [FREE Full text] [doi: 10.1093/nar/gkv1075] [Medline: 26481350]
41. Data files for predicting adverse drug-drug interactions with neural embedding of semantic predications. Zenodo. URL: https://zenodo.org/record/3333834/ [accessed 2021-09-24]
42. kb_embeddings. BitBucket. URL: https://bitbucket.org/cdal/kb_embeddings [accessed 2021-10-06]

Abbreviations

ADE: adverse drug event
AP@50: average precision at 50

AUC: area under the receiver operating characteristic curve
AUPRC: area under the precision-recall curve
ESP: Embedding of Semantic Predications
EU-ADR: Exploring and Understanding Adverse Drug Reactions
KG: knowledge graph
KNN: k-nearest neighbors
LOO: leave-one-out
LR: logistic regression
NLP: natural language processing
OMOP: Observational Medical Outcomes Partnership
RF: random forest
S5F: stratified 5-fold
SemMedDB: Semantic MEDLINE Database
SGNS: skip-gram negative sampling
SIDER: Side Effect Resource
TransE: Translating Embeddings for Modeling Multi-relational Data
t-SNE: t-distributed stochastic neighbor embedding
XOR: exclusive OR

Edited by G Eysenbach; submitted 08.08.21; peer-reviewed by X Dong; comments to author 27.08.21; revised version received 07.09.21; accepted 18.09.21; published 25.10.21.

Please cite as:
Dasgupta S, Jayagopal A, Jun Hong AL, Mariappan R, Rajan V
Adverse Drug Event Prediction Using Noisy Literature-Derived Knowledge Graphs: Algorithm Development and Validation
JMIR Med Inform 2021;9(10):e32730
URL: https://medinform.jmir.org/2021/10/e32730
doi:10.2196/32730
PMID:34694230

Original Paper

Predicting the Easiness and Complexity of English Health Materials for International Tertiary Students With Linguistically Enhanced Machine Learning Algorithms: Development and Validation Study

Wenxiu Xie[1], MSc; Christine Ji[2], PhD; Tianyong Hao[3], PhD; Chi-Yin Chow[1], PhD

[1]Department of Computer Science, City University of Hong Kong, Hong Kong, Hong Kong

[2]School of Languages and Cultures, University of Sydney, Sydney, Australia

[3]School of Computer Science, South China Normal University, Guangzhou, China

Corresponding Author:
Christine Ji, PhD
School of Languages and Cultures
University of Sydney
City Road Camperdown/Darlington
Sydney, 2006
Australia
Phone: 61 0434069975
Email: christine.ji@sydney.edu.au

Abstract

Background: There is an increasing body of research on the development of machine learning algorithms in the evaluation of online health educational resources for specific readerships. Machine learning algorithms are known for their lack of interpretability compared with statistics. Given their high predictive precision, improving the interpretability of these algorithms can help increase their applicability and replicability in health educational research and applied linguistics, as well as in the development and review of new health education resources for effective and accessible health education.

Objective: Our study aimed to develop a linguistically enriched machine learning model to predict binary outcomes of online English health educational resources in terms of their easiness and complexity for international tertiary students.

Methods: Logistic regression emerged as the best performing algorithm compared with support vector machine (SVM) (linear), SVM (radial basis function), random forest, and extreme gradient boosting on the transformed data set using L2 normalization. We applied recursive feature elimination with SVM to perform automatic feature selection. The automatically selected features (n=67) were then further streamlined through expert review. The finalized feature set of 22 semantic features achieved a similar area under the curve, sensitivity, specificity, and accuracy compared with the initial (n=115) and automatically selected feature sets (n=67). Logistic regression with the linguistically enhanced feature set (n=22) exhibited important stability and robustness on the training data of different sizes (20%, 40%, 60%, and 80%), and showed consistently high performance when compared with the other 4 algorithms (SVM [linear], SVM [radial basis function], random forest, and extreme gradient boosting).

Results: We identified semantic features (with positive regression coefficients) contributing to the prediction of easy-to-understand online health texts and semantic features (with negative regression coefficients) contributing to the prediction of hard-to-understand health materials for readers with nonnative English backgrounds. Language complexity was explained by lexical difficulty (rarity and medical terminology), verbs typical of medical discourse, and syntactic complexity. Language easiness of online health materials was associated with features such as common speech act verbs, personal pronouns, and familiar reasoning verbs. Successive permutation of features illustrated the interaction between these features and their impact on key performance indicators of the machine learning algorithms.

Conclusions: The new logistic regression model developed exhibited consistency, scalability, and, more importantly, interpretability based on existing health and linguistic research. It was found that low and high linguistic accessibilities of online health materials were explained by 2 sets of distinct semantic features. This revealed the inherent complexity of effective health communication beyond current readability analyses, which were limited to syntactic complexity and lexical difficulty.

(JMIR Med Inform 2021;9(10):e25110) doi:10.2196/25110

KEYWORDS

feature selection; logistic regression; online health resources

Introduction

For a long time, the study of the quality of language for effective health communication and education has focused on the complexity of health and medical educational resources [1-5]. A range of readability assessment tools have been developed to measure the lexical, grammatical, and syntactic features of health and educational resources [6-9]. Existing research shows that lack of linguistic understandability or readability can be explained by lexical difficulty, and complex grammatical and syntactic features [10-12]. This has caused the wide assumption that controlling for these textual features alone can help achieve the optimized reading experiences of medical and health materials for most people [13,14].

More recently, increasing research efforts have been geared toward developing accessible or easy-to-understand health materials and resources to help reduce the widening health inequality caused by socioeconomic determinants in societies with large and diverse vulnerable populations [15-19]. The key research question is whether previous studies and insights gained into health material readability can be translated directly into the design and development of accessible health resources for diverse populations, or whether there is a one-size-fits-all approach to accessible health information evaluation.

Natural language processing tools and machine learning algorithms have gained increasing popularity in health informatics. These flexible and versatile computational techniques can achieve high-precision prediction of outcomes based on the data-driven learning and computing of quantifiable features of the study object [1,7,10,20]. This represents a significant advance from statistics, which requires the presence of both dependent and independent variables to fit their relations into developed statistical models [21]. In the deployment stage, validated machine learning algorithms do not require the outcome variable to be available, as the algorithms can effectively predict the outcome, either a categorical or continuous variable, based on the computational learning of relevant features of the study object [22].

Providing reliable high-precision prediction of the outcome variable, for example, new online health information before release to the intended readers, can help identify and reduce potential barriers to health information understanding and thus increase the wide social accessibility of critical health information among diverse vulnerable populations or populations at risk due to lack of English proficiency and exposure to English health educational traditions.

Using first-hand materials from a diverse range of English health websites, our study developed a high-performing machine learning model to effectively predict the easiness versus difficulty of original English health information among young adults from nonnative English-speaking backgrounds. The machine learning algorithm revealed that while the difficulty of English health information can be explained by existing readability research, such as lexical unfamiliarity, medical terminology, and jargons, as well as syntactic complexity that can be measured by long and complex sentence structures, the easiness or understandability of original English material can be explained by distinct textual and semantic features associated with the use of common speech act verbs, familiar verbs of mental acts and processes (understand, learn, trust, feel, remember, etc), personal names and pronouns, names of social groups and communities, affiliations, people's relations, expressions that assist with the evaluation of events, scenarios, or circumstances such as probability expressions (can, might, may be, etc), purposeful expressions that direct or draw the attention of the readers to key points of the reading material such as adverbs describing levels, and degrees.

Methods

Data Collection

We collected 1000 original health texts published by national and international health authorities on a wide range of health topics ranging from infectious diseases, noncommunicable diseases (like cancers, diabetes, and cardiovascular diseases), and environmental health to mental diseases, disability, and palliative care. These materials were collected and screened for their information validity. We only kept health texts published by health authorities and organizations that have extensive experience in developing and disseminating credible health information [23-25]. Private, commercial, or nonaccredited health websites were excluded manually to ensure the reliability and usability of our research findings for the development, evaluation, and prediction of public-oriented health educational resources. The collected health materials were then divided into 2 categories of easy versus difficult materials by a small group of young adults with nonnative English-speaking backgrounds. They rated the texts on a continuous scale of 0 to 10, with 0 indicating the easiest level and 10 indicating the hardest level. Their original ratings were then standardized to z scores, and the mean z score was taken as the notional value of the reading difficulty of a certain text. Lastly, we took the grand mean of the z scores of the 1000 texts and classified the entire corpus with a binary classification framework. Those below the grand mean were labeled as easy to read texts, and those above the grand mean were labeled as difficult health texts for machine learning algorithm development.

Machine Learning Algorithm Development

Data Normalization

The 1000 health education articles were randomly split into training data and test data with 700 and 300 samples, respectively. The training data were used for 5-fold cross-validation to select the best hyperparameters for the machine learning algorithms, and the test data were used for evaluation and validation. The statistic distribution of the

training and test data were as follows: training data, 384 difficult and 316 easy samples; test data, 162 difficult and 138 easy samples. Data normalization is necessary and essential for the machine learning algorithms to achieve good generalization and classification performance [26,27], which normalizes and scales the range of data features to prevent those features with a larger range from dominating the optimization process. It can also improve the learning process, and the range of the value of the collected health text data varies widely from a minimal value of 0 to a maximal value of 1030. Thus, we performed the following 3 normalization methods implemented in scikit-learn [28] on the data: min-max normalization, z-score normalization, and L_2-norm normalization. Min-max normalization scales the data to a certain range like (0,1) or (−1,1). For z-score normalization, the rescaled data would have a unit variance and 0 mean. The L_2-norm normalization scales the data samples individually to unit norm, and the sum of the squares of the data will always be up to 1. The formulas of these 3 methods are shown in equations (1), (2), and (3), respectively. For the data sample x, the minimum value is denoted as x_{min} and the maximum value is denoted as x_{max}. The mean of the data is denoted as x_{mean}, and the SD of the data is denoted as x_{std}.

Hyperparameter Tuning and Model Selection

We evaluated the following 5 machine learning algorithms on our constructed health education data: linear model logistic regression (LR), linear model support vector machine (SVM) with linear kernel, nonlinear model SVM with radial basis function (RBF) kernel, ensemble tree model random forest (RF), and extreme gradient boosting (XGBoost) [29]. The nonlinear and ensemble tree models were able to learn a decision boundary that is nonlinear in the input space. The algorithms were implemented in Python with scikit-learn and xgboost packages.

To optimize the performance of the machine learning algorithms, we performed leave-one-out 5-fold cross-validation on the training data to fine tune the hyperparameters of each model via automatic grid search and randomized search methods. For LR, SVM (linear), and SVM (RBF), where the candidate values of hyperparameters are discrete, we applied a grid search to perform an exhaustive search to find the best and cross-validated parameter values of the model. For the ensemble tree model RF and XGBoost, where some of the hyperparameters are continuous, the randomized search method was applied to save the hyperparameter tuning space and time, which sampled a fixed number of parameter settings from the specified distribution instead of performing an exhaustive search. Figures 1 and 2 show the hyperparameter tuning process of the LR and SVM (linear) models, respectively. The fine-tuned values of core hyperparameters of the models are shown in Table 1. For hyperparameters not listed, we used the default value in the model.

Figure 1. Hyperparameter tuning process of logistic regression. CV: cross-validation.

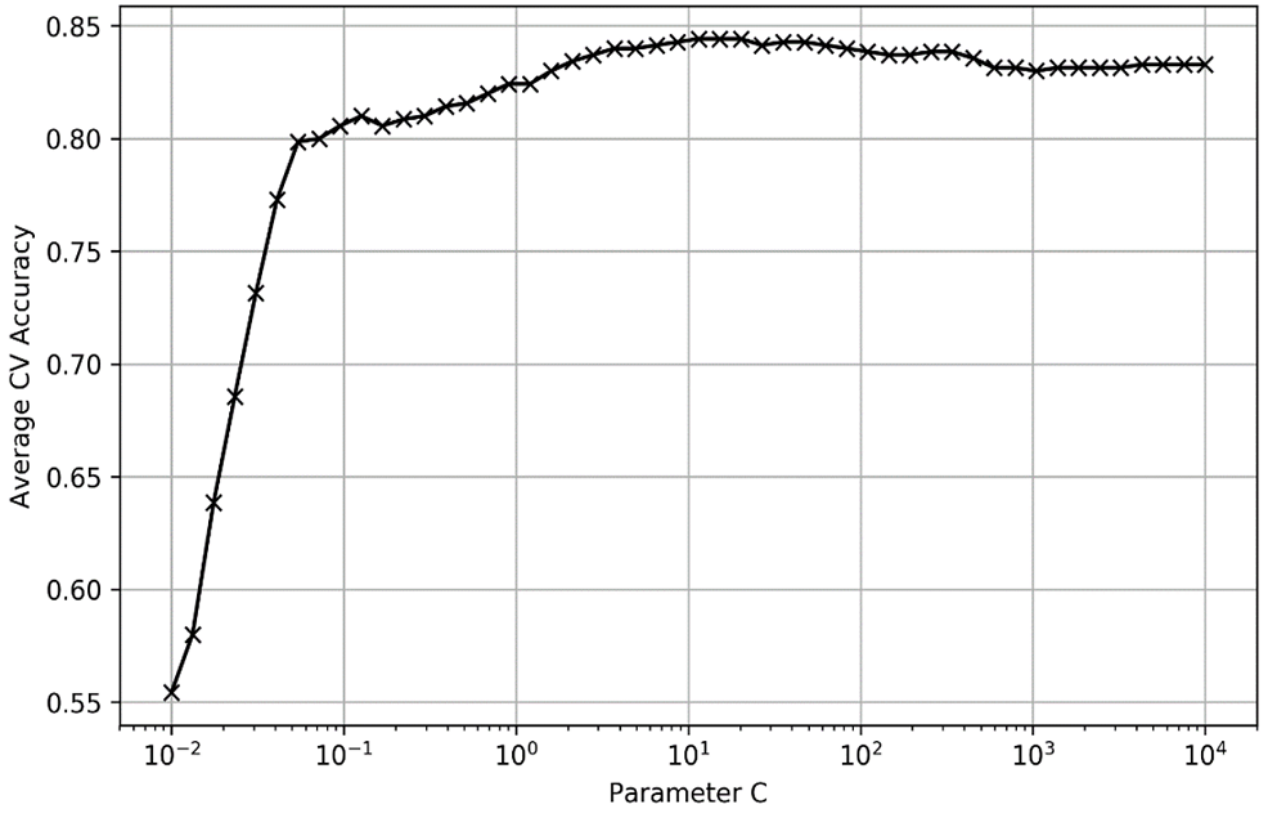


Figure 2. Hyperparameter tuning process of support vector machine (linear). CV: cross-validation.

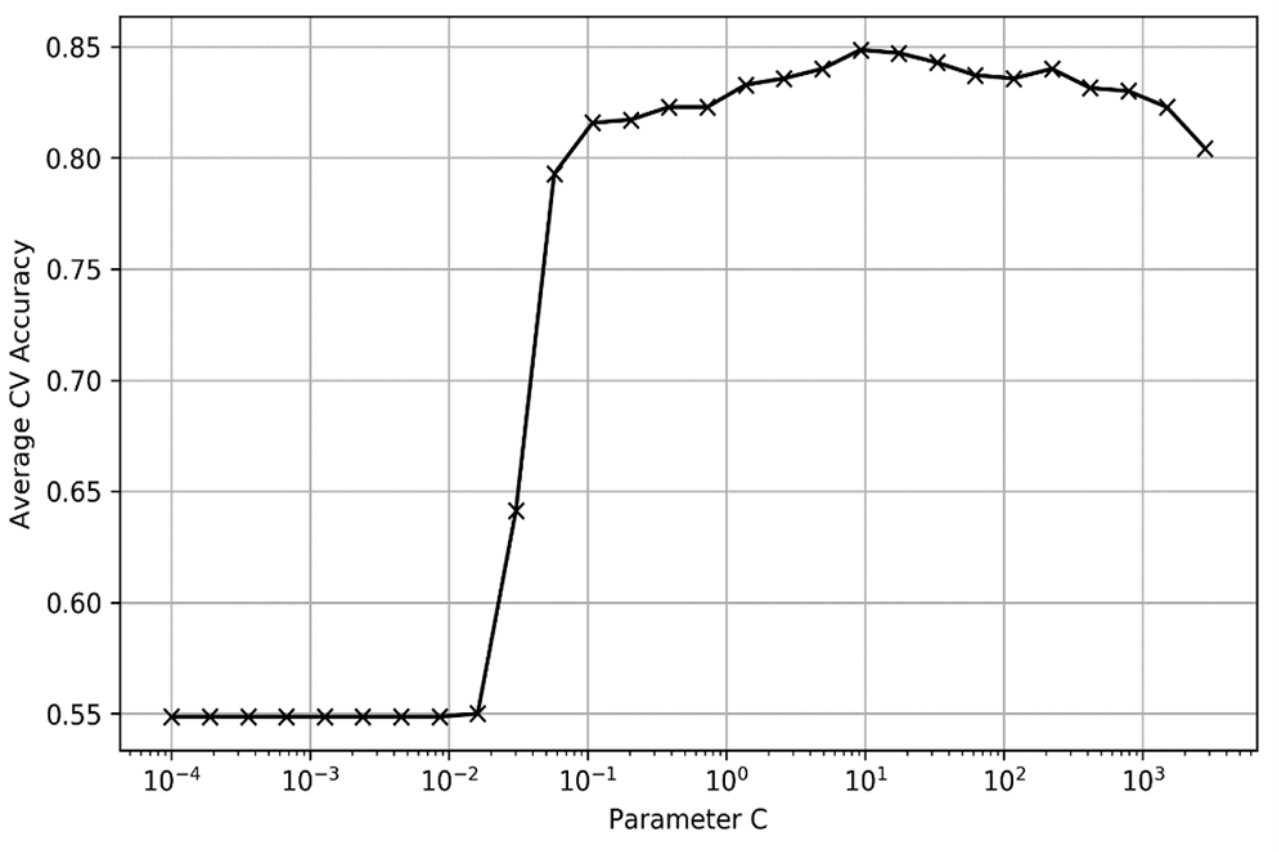


Table 1. The fine-tuned values of the core hyperparameters of machine learning algorithms.

Algorithm and hyperparameter name	Description	Value
Logistic regression		
C	Inverse of regularization strength	11.51395399
Support vector machine (linear)		
C	Regularization parameter. The strength of the regularization is inversely proportional to C.	9.236708571873866
kernel	The kernel type to be used in the algorithm.	linear
probability	Whether to enable probability estimates.	True
Support vector machine (radial basis function)		
C	Regularization parameter. The strength of the regularization is inversely proportional to C.	221.22162910704503
gamma	Kernel coefficient.	0.02212216291070450
kernel	The kernel type to be used in the algorithm.	rbf
probability	Whether to enable probability estimates.	True
Extreme gradient boosting (XGBoost)		
subsample	Subsample ratio of the training instances.	0.7842105263157895
n_estimators	Number of boosted trees to fit.	120
min_child_weight	Minimum sum of instance weight (hessian) needed in a child.	2
max_depth	Maximum depth of a tree.	4
learning_rate	Step size shrinkage used in an update to prevent overfitting.	0.11473684210526315
colsample_bytree	The subsample ratio of columns when constructing each tree.	0.5666666666666667
RFE_SVM[a]		
estimator	A supervised learning estimator with a fit method that provides information about feature importance.	SVM
min_features_to_select	The minimum number of features to be selected.	1
step	The number of features to remove at each iteration.	1
Random forest		
n_estimators	The number of trees in the forest	100
max_depth	The maximum depth of the tree.	4
min_samples_leaf	The minimum number of samples required to split an internal node.	0.0463703639292683
min_samples_split	The minimum number of samples required to be at a leaf node.	0.06216133047419098

[a]RFE_SVM: recursive feature elimination_support vector machine.

Results

Statistical Analyses

We annotated the corpus with the semantic annotation system developed by Lancaster University, UK [30]. The features were count data. Table 2 shows the results of the Mann-Whitney *U* test of the 2 sets of health texts across 22 of the original 115 semantic features as an illustration of contrasts between easy and difficult texts. Statistically significant differences ($P<.05$) existed for most features. To help with the understanding of the annotated semantic features, some typical words of each feature were extracted from the original corpus. A2 included cause, affect, trigger, develop, progression, depend, evoke, transmission, modify, etc. Examples from texts classified as difficult are “neurodegenerative condition that *affects* the central nervous system,” “in some cases, *evoked* potentials (nerve transmission speed) may be measured and/or a lumbar puncture (spinal tap) may be required,” and “an attack *results in* inflammation and *development* of one or more lesions, resulting in scarring (sclerotic plaque), forming on the nerves.” A7 included can, may, might, could, must, etc. Examples from texts classified as easy to understand are “the second section has an orange border and *can* help you understand what *may* have happened to you,” “you *can* ask someone you trust to help you with these books. This *might* be a disability support worker or a family violence support worker,” and “these books are about where violence *can* happen and who *can* do violence.” A12 included problems, hard, tough, etc. Examples from texts classified as easy to understand are “in fact, most kids run away due to *problems* with their families,” “anger is one of the hardest

emotions to manage because it's so strong," "if your friend is thinking about running away, warn him or her about how tough it will be to survive on the streets." A13 included most, at least, thoroughly, etc. Examples from texts classified as easy to understand are "thoroughly wash your hands beforehand to reduce the risk of spreading the infection to others." A15 included risk, safe, dangerous, danger, exposure, at risk, etc. B1 included heart, muscles, chest, blood vessel, artery, ventricle, valve, cardiovascular, and contraction. B2 included arrhythmia, arteriosclerosis, abnormalities, and Down syndrome. B3 included oxygen mask, medication, antibiotics, vaccines, diagnosis, paracetamol, ibuprofen, steroids, etc. Q2 included admit, deliver, talk, speak, call, acknowledge, advice, suggest, note, question, answer, voice, etc. S2 included women, girls, children, people, workers, staff, providers, etc. S3 included partners, friends, girlfriend, boyfriend, etc. S5 included member, organization, public, community, board, group, personal, etc. T2 included begin, start to, still be, hold off on, go on and on, get going; during, all the time, end with, etc. X2 included understand, learn, trust, feel, remember, seek, check, experience, reason, inform, review, etc. X7 included want, purposes, planning, mission, aim, target, requirement, focus, etc. Y1 included radiation, x-rays, telescope, bioterrorism, tissue engineering, anatomy, laser, etc. Y2 included software, online, internet, email, computer, websites, screen, etc. Z6 included not, no, and negative. Z99 included paratyphoid, bleach based, handwashing, alcohol based, ready to eat, cross-contamination, unpasteurized, disinfect, salmonellosis, cardiologists, parainfluenza, etc.

Table 2. Mann-Whitney *U* test results.

Code	Definition	Easy; mean (SD)	Difficult; mean (SD)	Asymptotic significance (2-tailed) of the mean difference (*P* value)
A2	Cause and effect	9.03 (9.41)	10.64 (12.83)	.12
A7	Probability	8.55 (10.48)	4.18 (7.30)	<.001
A12	Easy/difficult	1.54 (2.35)	0.73 (1.55)	<.001
A13	Degree descriptors	4.75 (5.57)	3.96 (4.63)	.07
A15	Safety/danger	1.17 (3.37)	1.46 (3.87)	.83
B1	Anatomy and physiology	17.09 (31.84)	15.80 (21.80)	.16
B2	Health and disease	14.66 (21.20)	24.27 (33.45)	<.001
B3	Medicines and medical treatment	8.97 (14.01)	12.77 (17.93)	<.001
Q2	Speech acts	9.67 (11.76)	5.68 (8.57)	<.001
S2	People	12.17 (16.58)	9.31 (16.12)	<.001
S3	Relationship	1.72 (4.48)	0.79 (3.14)	<.001
S5	Groups and affiliation	3.05 (5.68)	1.75 (3.38)	<.001
T2	Time	2.95 (4.09)	2.41 (3.35)	.07
X2	Mental actions and processes	10.69 (11.19)	6.62 (9.35)	<.001
X7	Intention/purposes	1.78 (3.13)	3.39 (5.55)	<.001
Y1	Science and technology in general	0.19 (0.58)	0.73 (1.52)	<.001
Y2	Information technology and computing	1.33 (3.34)	0.60 (2.09)	<.001
Z1	Personal names	2.06 (4.16)	3.05 (6.90)	.51
Z5	grammatical expressions	120.84 (120.48)	136.93 (116.84)	<.001
Z6	Negative	3.01 (5.11)	4.23 (5.17)	<.001
Z8	Pronouns	54.97 (48.37)	21.14 (30.53)	<.001
Z99	Unmatched expressions	18.63 (29.12)	45.70 (54.20)	<.001

Model Selection

After hyperparameter tuning, we compared the performance of 5 machine learning models with different normalization methods to select the best model for our further feature selection/reduction validation. We reported 5-fold cross-validation average accuracy on the training data and accuracy on the test data of all models. For 5-fold cross-validation, we applied a different random seed so the 5-fold data sets for validation would be different from the data sets for hyperparameter tuning. The results are shown in Table 3. As shown in the results, L_2-norm normalization can improve the classification performance of the LR, SVM (linear), and SVM (RBF) models on both training data and test data. However, z-score and min-max normalization have negative impacts on model performance. For the ensemble tree model RF and XGBoost, data normalization is unnecessary in model development since the large range of feature values can help the partitioning process. LR with L_2-norm normalization, which

yielded the best performance on both training data and test data, was selected as the best model for further validation.

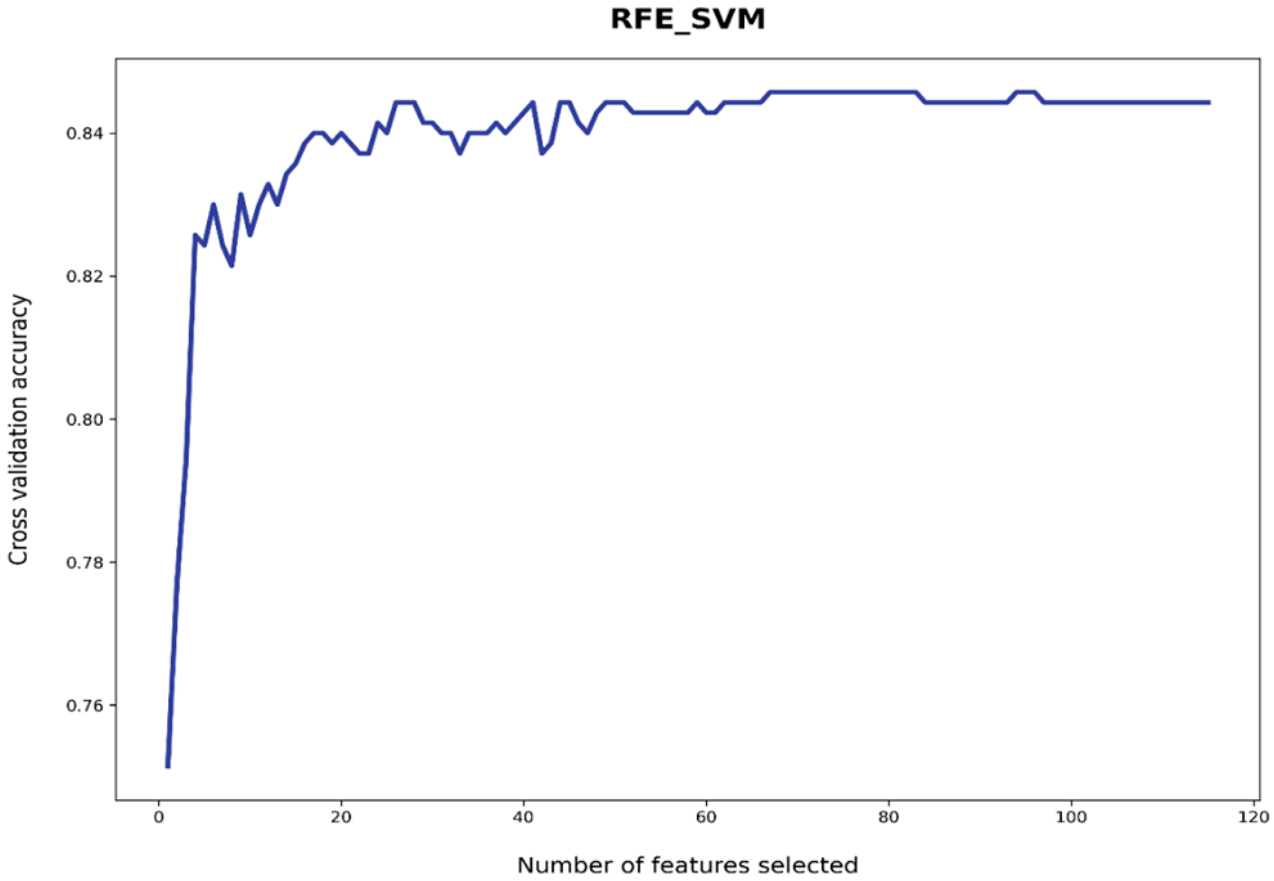
Table 3. Performance of the 5 selected models with different data normalization methods.

Classifier	Not normalized		Z-score		Min-Max		L_2-norm	
	Training	Test	Training	Test	Training	Test	Training	Test
Logistic regression	0.823	0.817	0.810	0.777	0.809	0.780	0.840	0.840
Support vector machine (linear)	0.816	0.823	0.819	0.803	0.801	0.807	0.833	0.840
Support vector machine (radial basis function)	0.833	0.823	0.807	0.793	0.829	0.787	0.834	0.833
Random forest	0.796	0.803	0.796	0.803	0.796	0.803	0.789	0.830
Extreme gradient boosting	0.821	0.840	0.821	0.840	0.837	0.837	0.803	0.837

Automatic Feature Selection: 67 Features

We applied recursive feature elimination (RFE) with SVM as the base estimator (RFE_SVM) to learn feature importance and performed feature reduction to remove unimportant features [31]. During the feature selection process, RFE_SVM decides whether a certain selected feature is useful or not for the SVM model to learn the decision boundary. This was achieved via iteratively eliminating features. Figure 3 shows the automatic cross-validated tuning process of RFE_SVM with different numbers of selected features. RFE_SVM learned 67 features, eliminating 48 unimportant features from the original full feature set of 115 features. The learned 67 features were automatically selected (AS) features from machine learning algorithms.

Figure 3. Automatic cross-validated tuning results of recursive feature elimination_support vector machine (RFE_SVM) with different numbers of selected features.



Expert Feature Review and Refinement

Out of the 67 features selected automatically by the machine learning algorithm (RFE_SVM), 45 features were manually eliminated in the following expert review, as these features were mostly words/expressions that were not directly relevant to health or medical information. These included A1 (general actions); A5 (evaluation [good/bad, true/false]); A6 (comparison [similar/different]); A9 (getting and giving [possession]); A10 (open, finding, showing); B4 (cleaning and personal care); C1 (arts and crafts); E3 (calm/violent/angry emotions); E4 (happiness and contentment emotions); F1 (food); F2 (drinks and alcohol); F3 (smoking and nonmedical drugs); F4 (farming and horticulture); G1 (government and politics); G2 (crime, law

and order); H5 (furniture and household fittings); I1 (money generally); L1 (life and living things); L2 (living creatures); L3 (plants); M1 (moving, coming, and going); M3 (vehicles, transport on land); M5 (flying/aircraft); M6 (location and direction); M8 (stationary); O2 (objects generally); O4 (physical attributes); P1 (education in general); S1 (social actions, states, processes); S7 (social actions, states, processes); S9 (religion, supernatural); T1 (time); W3 (geographical terms); and so on. The automatic feature selection reduced the original features by 41.7% from 115 to 67 features, and the subsequent expert review reduced a further 39.1% from 67 to 22 features. Tables 4 and 5 show the comparison of the performance of LR selected as the best performing algorithm with the following 3 sets of features: 115, 67, and 22. We used 70% of the data set as training data and 30% as test data, and then applied 3-fold cross-validation. The pair-wise corrected resample *t* test showed that with a significantly reduced number of features, the performance of the algorithm was not affected, even with a slightly better improvement in terms of model accuracy (mean difference of accuracy between 22 and 67 features: *P*=.04).

Table 4. Performance of machine learning models using different sets of features as predictors (logistic regression).

Algorithm	Accuracy, mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)	AUC[a], mean (SD)	Macro F1, mean (SD)
115 features	0.8400 (0.0100)	0.7767 (0.0321)	0.8933 (0.0416)	0.9188 (0.0048)	0.8367 (0.0153)
67 features	0.8400 (0.0200)	0.8333 (0.0945)	0.8333 (0.0643)	0.9177 (0.0066)	0.8367 (0.0252)
22 features	0.8567 (0.0208)	0.8100 (0.0173)	0.8933 (0.0503)	0.9108 (0.0045)	0.8567 (0.0208)

[a]AUC: area under the curve.

Table 5. Pair-wise corrected resampled *t* test of accuracy differences (using 3 sets of features as predictors).

Comparison	Mean difference	95% CI		*P* value (2-tailed)
		Lower	Upper	
Pairwise comparison of accuracy				
Pair 1: 67 features vs 115 features	0.00%	−0.0196	0.0196	>.99
Pair 2: 22 features vs 115 features	1.98%	−0.0060	0.0393	.13
Pair 3: 22 features vs 67 features	1.98%	0.0054	0.0280	.04
Pairwise comparison of sensitivity				
Pair 1: 67 features vs 115 features	7.30%	−0.1884	0.3017	.52
Pair 2: 22 features vs 115 features	4.29%	−0.0265	0.0932	.20
Pair 3: 22 features vs 67 features	−2.80%	−0.2329	0.1862	.74
Pairwise comparison of specificity				
Pair 1: 67 features vs 115 features	−6.72%	−0.2637	0.1437	.42
Pair 2: 22 features vs 115 features	0.00%	−0.0392	0.0392	>.99
Pair 3: 22 features vs 67 features	7.20%	−0.1474	0.2674	.43
Pairwise comparison of AUC[a]				
Pair 1: 67 features vs 115 features	−0.12%	−0.0079	0.0056	.63
Pair 2: 22 features vs 115 features	−0.87%	−0.0264	0.0103	.28
Pair 3: 22 features vs 67 features	−0.75%	−0.0280	0.0142	.38
Pairwise comparison of F1				
Pair 1: 67 features vs 115 features	0.00%	−0.0196	0.0196	>.99
Pair 2: 22 features vs 115 features	2.39%	0.0004	0.0396	.07
Pair 3: 22 features vs 67 features	2.39%	0.0004	0.0396	.07

[a]AUC: area under the curve.

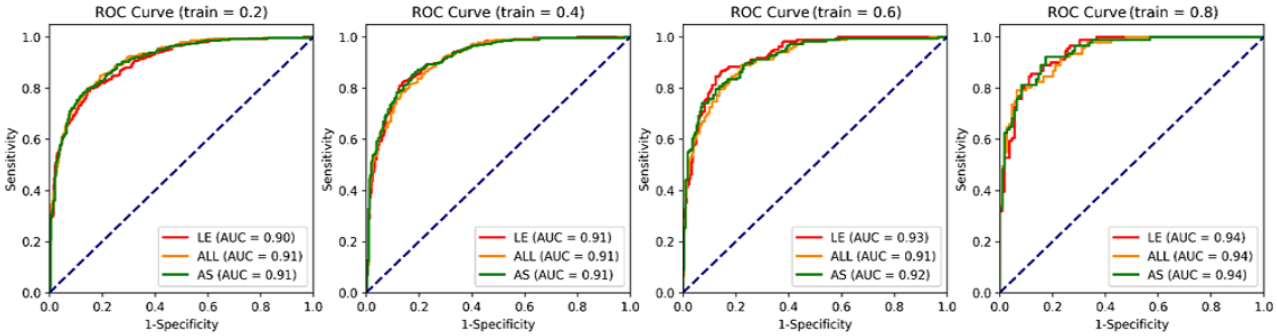
Model Validation

We evaluated the stability, robustness, scalability, and effectiveness of the 22 linguistically enhanced (LE) features. We first compared the performance of the LE features with all initial (ALL) 115 features and the 67 AS features on different sizes of training and test data. The entire data were randomly split into training data and test data with different split rates (0.2, 0.4, 0.6, and 0.8). For instance, with a split rate of 0.2 (denoted as train=0.2), 20% of data were used as training data and the remaining 80% of data were used as test data for validation. The receiver operating characteristic curve and area

under the curve (AUC) metrics were used to evaluate the model performance.

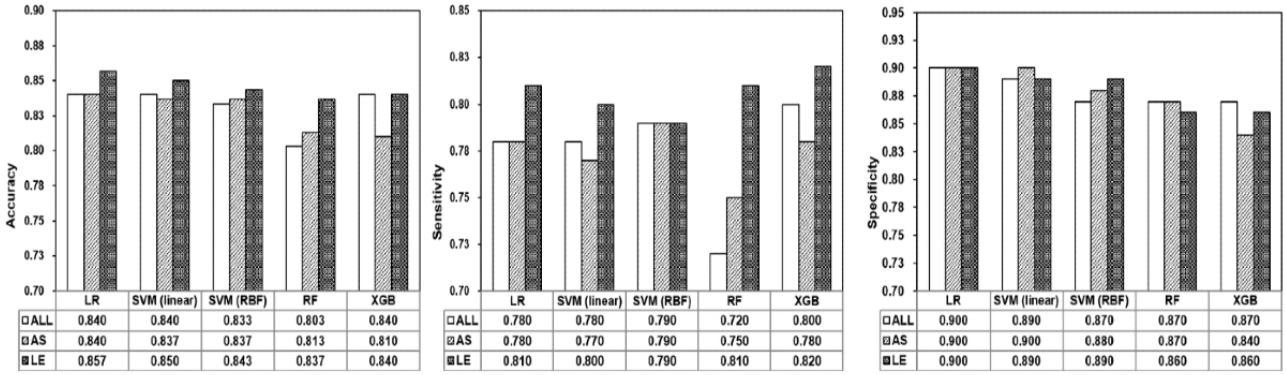
As shown in Figure 4, the model using LE features consistently yielded a comparable or higher performance on different training data set sizes (train=0.4, 0.6, and 0.8) compared with the models involving ALL and AS features. For the split rate of 0.2, the model with LE features had a lower AUC score compared with the models involving ALL and AS features. This was caused by the underfitting nature of the model involving LE features (with only 22 features, less variance, and more bias). The models involving ALL and AS features were more likely to be overfitting for the number of training data (n=200), which was very close to the number of features used for classification (115 and 67, respectively). With an increase in the training data size, the underfitting issue was solved, and the model involving LE features had better performance compared with the models involving ALL and AS features.

Figure 4. Stability and robustness of the 22 linguistically enhanced features. ALL: all initial (115 features); AS: automatically selected (67 features); AUC: area under the curve; LE: linguistically enhanced (22 features); ROC: receiver operating characteristic.



To better evaluate the scalability, effectiveness, and contribution of the constructed LE features on health educational material classification, we compared the performance of 5 machine learning models with different feature sets. The selected machine learning models were LR, SVM (linear), SVM (RBF), RF, and XGBoost, and the models were trained on training data and evaluated on test data. The performance was assessed in terms of accuracy, sensitivity, and specificity metrics. As shown in Figure 5, LE features benefited all machine learning models compared with AS and ALL features, with a comparable or higher accuracy and sensitivity. The models (RF and XGBoost) with LE features had lower specificity than the models with AS and ALL features. Overall, LE features had a positive impact on the machine learning process, and changing the machine learning model will not affect the overall learning performance on health education data, demonstrating its scalability and effectiveness.

Figure 5. Scalability and effectiveness of the 22 linguistically enhanced features. ALL: all initial (115 features); AS: automatically selected (67 features); LE: linguistically enhanced (22 features); LR: logistic regression; RBF: radial basis function; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting.



Impact of Features on Model Sensitivity and Specificity

The final feature set contained 22 features, as presented in Figure 6, which shows the regression coefficients of the finalized 22 features. Half of the features (n=11) were associated with the easiness of the health materials, as indicated by their positive regression coefficients as follows: Z8 (pronouns), 5.717293; S5 (groups/affiliations), 3.393270; X2 (mental actions and processes), 3.350851; A7 (probability), 2.942145; A12 (easy versus difficult), 2.610647; A13 (degree descriptors), 1.462447; Q2 (speech verbs), 1.093459; Y2 (information technology/computing), 0.949234; S3 (relations), 0.898446; Z1 (personal names), 0.548855; and S2 (people), 0.135449. The other half of the features (n=11) were associated with the difficulty of the health materials, as indicated by their negative regression coefficients as follows: Z6 (negative functional words), −0.221485; X7 (intentions), −0.743811; Y1 (science and technology), −1.903669; A15 (safety/risks), −2.291032; T2 (time), −2.756571; B1 (anatomy and physiology), −3.021697; B2 (health and disease), −3.793444; A2 (cause and effect verbs), −4.838672; B3 (medicines and medical treatment), −5.763809;

Z5 (grammatical expressions), −7.348969; and Z99 (unmatched or out-of-dictionary expressions), −8.749430.

Figure 6. Feature coefficients in the logistic regression model with 22 semantic features. Descriptors are shown in Multimedia Appendix 1.

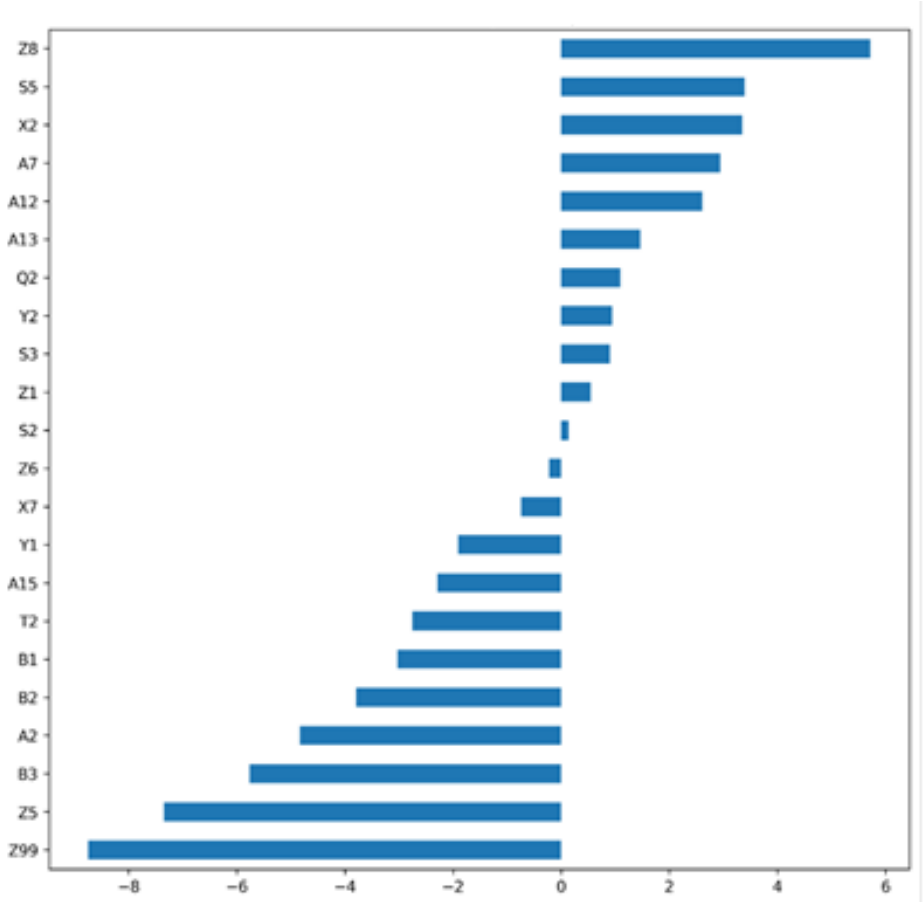
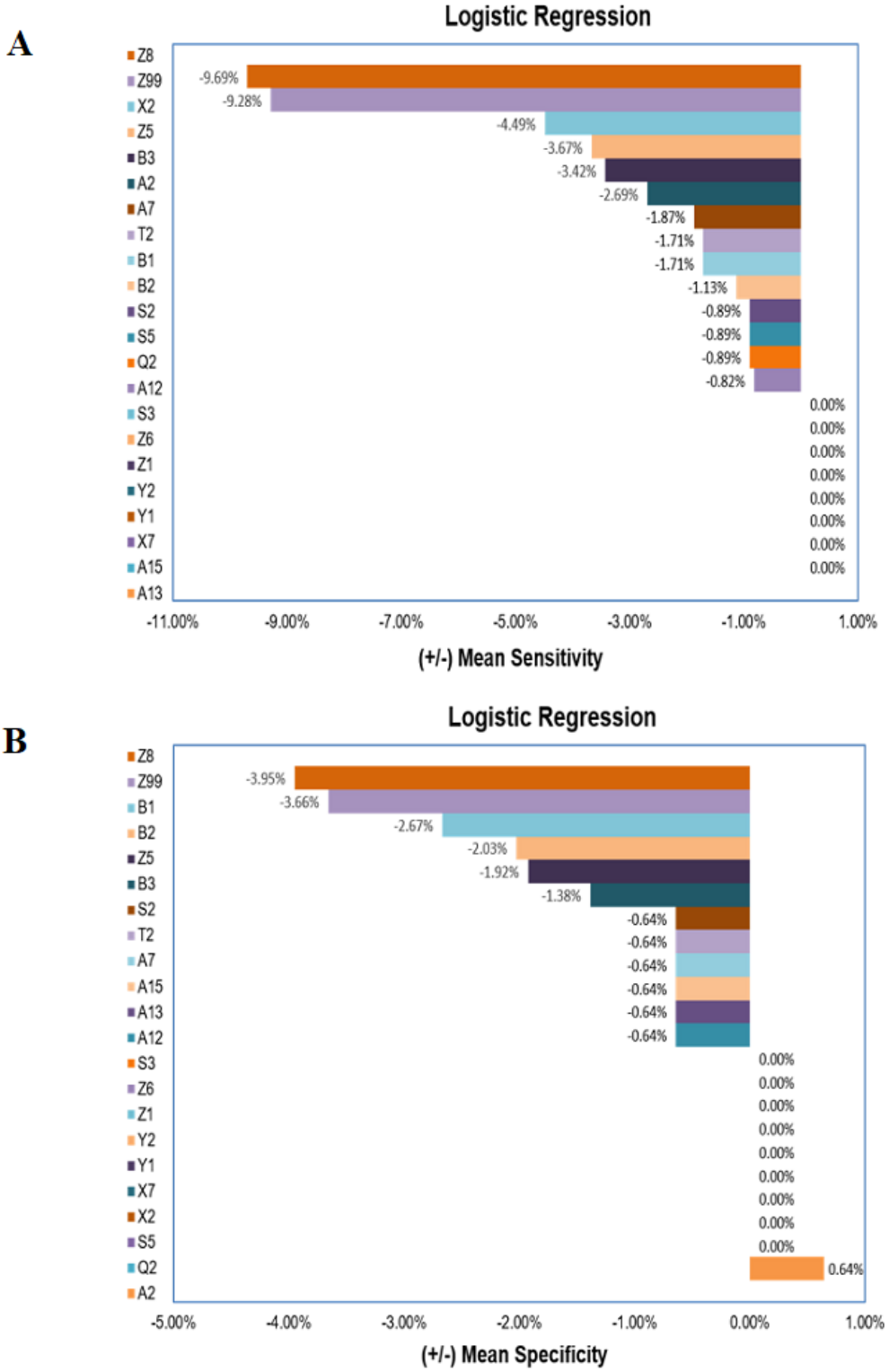


Figure 7A shows the impact of individual features on the sensitivity of the LR algorithm. The removal of Z8 (pronouns) resulted in a mean decrease in sensitivity of 9.69% (*P*=.047; 95% CI of mean difference −0.1387 to −0.0189). The removal of A2 (cause and effect) reduced the model sensitivity by 2.69% to a mean sensitivity score of 0.7907 (SD 0.0163; *P*=.003; 95% CI −0.0256 to −0.0181). Another feature that caused a statistically significant drop in model sensitivity was X2 (mental actions and processes). The deletion of this feature resulted in a mean sensitivity decrease of 4.49%, from a mean sensitivity score of 0.8126 to 0.7761 (SD 0.014; *P*=.04; 95% CI −0.0625 to −0.0104). The following 11 features also caused decreases in the model mean sensitivity, but the changes were not statistically significant: A12 (easy/difficult; mean difference −0.82%, change to 0.8059; *P*=.42; 95% CI −0.0293 to 0.016); A7 (probability; mean difference −1.87%, change to 0.7974; *P*=.18; 95% CI −0.041 to 0.0107); B1 (anatomy and physiology; mean difference −1.71%, change to 0.7986; *P*=.18; 95% CI −0.0376 to 0.0098); B2 (health and diseases; mean difference −1.13%, change to 0.8034; *P*=.69; 95% CI −0.0773 to 0.0589); B3 (medicines and medical treatment; mean difference −3.42%, change to 0.7847; *P*=.18; 95% CI −0.0752 to 0.0195); Q2 (speech act verbs; mean difference −0.89%, change to 0.8053; *P*=.42; 95% CI −0.0318 to 0.0174); S5 (groups and affiliations; mean difference −0.89%, change to 0.8053; *P*=.42; 95% CI −0.0318 to 0.0174); Z5 (grammatical expressions; mean difference −3.67%, change to 0.7828; *P*=.08; 95% CI −0.0601 to 0.0005); Z99 (unmatched/out-of-dictionary words; mean difference −9.28%, change to 0.7371; *P*=.16; 95% CI −0.1899 to 0.039); T2 (time; mean difference −1.71%, change to 0.7986; *P*=.18; 95% CI −0.0376 to 0.0098); and S2 (people; mean difference −0.89%, change to 0.8053; *P*=.42; 95% CI −0.0318 to 0.0174). Figure 7B shows the impact of features on the specificity of the LR model. Decreases in specificity with removal were noted for the following features: A12 (easy/difficult; mean difference −0.64%, change to 0.8907; *P*=.42; 95% CI −0.0253 to 0.0138); A13 (degree descriptors; mean difference −0.64%, change to 0.8907; *P*=.42; 95% CI −0.0253 to 0.0138); A15 (safety/risks; mean difference −0.64%, change to 0.8907; *P*=.42; 95% CI −0.0253 to 0.0138); A7 (probability; mean difference −0.64%, change to 0.8907; *P*=.42; 95% CI −0.0253 to 0.0138); B1 (anatomy and physiology; mean difference −2.67%, change to 0.8725; *P*=.25; 95% CI −0.075 to 0.0272); B2 (health and diseases; mean difference −2.03%, change to 0.8783; *P*=.21; 95% CI −0.0521 to 0.0158); B3 (medicine and medical treatments; mean difference −1.38%, change to 0.884, *P*=.19; 95% CI −0.0337 to 0.0088); Z5 (grammatical expressions; mean difference −1.92%, change to 0.8792; *P*=.42; 95% CI −0.0758 to 0.0413); Z8 (pronouns; mean difference −3.95%, change to 0.861; *P*=.31; 95% CI −0.1238 to 0.053); Z99 (unmatched expressions; mean difference −3.66%, change to 0.8637; *P*=.32; 95% CI −0.1178 to 0.0522); T2 (time; mean difference −0.64%, change to 0.8907; *P*=.42; 95% CI −0.0253 to 0.0138); and S2 (people; mean difference

−0.64%, change to 0.8907; *P*=.42; 95% CI −0.0253 to 0.0138). The only feature that caused an increase in specificity with its removal was A2 (cause and effect; mean difference 0.64%, change to 0.9022; *P*=.42; 95% CI −0.0138 to 0.0253).

Figure 7. Impact of the features on model sensitivity (A) and specificity (B). Descriptors are shown in Multimedia Appendix 1.



Discussion

Principal Findings

Improving the readability and accessibility of online English health education resources can have important impacts on the development of health literacy and the self-health management skills of readers. Young adults represent a large and increasing group of online health information consumers. Our study developed machine learning algorithms to predict the linguistic easiness versus difficulty for international tertiary students with non-English speaking backgrounds. We first compared and selected algorithms through data normalization (L_2-norm). LR

emerged as the best performing algorithm compared with SVM, RF, and XGBoost when trained on normalized data. We used RFE with SVM as the base estimator to automatically reduce the high-dimensional feature space. The automatic feature selection reduced the original feature set (n=115) by around 40% to 67.

The subsequent expert evaluation resulted in another 40% reduction in features. The distribution of regression coefficients aligns well with the statistical analyses. The following features with positive regression coefficients in machine learning had significantly higher means in easy-to-understand health materials (Mann-Whitney *U* test): Z8 (*P*<.001), S5 (*P*<.001), X2 (*P*<.001), A7 (*P*<.001), A12 (*P*<.001), Q2 (*P*<.001), Y2 (*P*<.001), S3 (*P*<.001), and S2 (*P*<.001). only the following 2 semantic features had statistically similar means in easy and difficult texts, with positive regression coefficients: A13 (*P*=.07) and Z1 (*P*=.51). The following features with negative regression coefficients had statistically higher means in difficult health materials: Z99 (*P*<.001), Z5 (*P*<.001), B3 (*P*<.001), B2 (*P*<.001), Y1 (*P*<.001), X7 (*P*<.001), and Z6 (*P*<.001). The following 4 semantic features had a statistically similar distribution in easy and difficult texts, with negative regression coefficients: A2 (*P*=.12), A15 (*P*=.83), B1 (*P*=.16), and T2 (*P*=.07). This suggests that statistical significance is not the only determinant in the development of LR algorithms. Feature interaction may also impact the performance of algorithms, although the impact of individual features on model sensitivity and specificity was not statistically significant.

To assess the impact of features on model performance, we conducted successive permutation of features to examine changes in the sensitivity and specificity of the LR algorithm. The LR model with the 22 optimized features achieved the highest sensitivity (mean 0.813, SD 0.018) and the highest specificity (mean 0.896, SD 0.050), when compared with other feature sets, which were optimized either automatically or statistically. Within the best performing model, the following 3 semantic features caused a statistically significant decrease in model sensitivity for predicting the linguistic easiness of online health information: Z8 (pronouns), A2 (words describing causes, effects, or causal relations), and X2 (words describing mental status, actions, or processes). We interpreted this important finding in light of the impact of an information logic sequence on reading experiences. The use of pronouns and words describing the causal relations can significantly increase the explicitness of the logical structure of health information [32-34]. The addition of words describing mental status, actions, and processes can help with the reasoning and mental processing of health information [35-38]. Different from the impact on sensitivity, none of the individual features caused a statistically significant decrease in specificity for predicting the difficulty of health materials for international tertiary students. This finding correlates well with existing research on readability. Linguistic features, such as word length, word frequency, and word familiarity, and other structural features have proven to be highly relevant and reliable predictors of textual complexity and difficulty [39].

Limitations

Our study developed an LR algorithm with a small number of features to predict the easiness and difficulty of online health information. The intended users were young adults with university degrees but with nonnative English-speaking backgrounds. The model is limited to this user group. The extensibility of our study findings to other user groups and online health materials in other languages remains to be tested and validated. Another limitation of our study is that the LR model using the finalized 22 features did not achieve statistically significant improvement over the LR model using the 115 and 67 features identified automatically. In future research, we will explore models that can achieve better performance with a small set of features that are linguistically meaningful and significant as well.

Conclusion

We developed a high-performing LR algorithm with a small number of semantic features to predict the easiness versus difficulty of online English health resources for young adults (tertiary students) with nonnative English-speaking backgrounds. We found that reducing the number of features is essential to prevent overfitting, since models with less features are less likely to have overfitting issues. Furthermore, machine learning models with less features are less complex, are more interpretable, and have better generalization [40]. The result also demonstrates the stability and robustness of the algorithm with linguistically relevant features. Our study shows that incorporating linguistic knowledge and machine learning–aided feature selection to reduce the feature space can help develop more efficient and less complex models with a good generalization ability.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Notations and definitions of semantic features.
[DOCX File , 20 KB - medinform_v9i10e25110_app1.docx]

References

1. Ji M, Liu Y, Zhao M, Lyu Z, Zhang B, Luo X, et al. Use of Machine Learning Algorithms to Predict the Understandability of Health Education Materials: Development and Evaluation Study. JMIR Med Inform 2021 May 06;9(5):e28413 [FREE Full text] [doi: 10.2196/28413] [Medline: 33955834]

2. Griffin J, McKenna K, Tooth L. Written health education materials: Making them more effective. Aust Occ Ther J 2003 Sep;50(3):170-177. [doi: 10.1046/j.1440-1630.2003.00381.x]
3. Demir F, Ozsaker E, Ilce A. The quality and suitability of written educational materials for patients. J Clin Nurs 2008 Jan;17(2):259-265. [doi: 10.1111/j.1365-2702.2007.02044.x] [Medline: 18171395]
4. Eysenbach G. Issues in evaluating health websites in an Internet-based randomized controlled trial. J Med Internet Res 2002 Dec 17;4(3):E17 [FREE Full text] [doi: 10.2196/jmir.4.3.e17] [Medline: 12554548]
5. French KS, Larrabee JH. Relationships among educational material readability, client literacy, perceived beneficence, and perceived quality. J Nurs Care Qual 1999 Aug;13(6):68-82. [doi: 10.1097/00001786-199908000-00008] [Medline: 10476626]
6. Gunning R. Technique of clear writing. New York, NY: McGraw Hill Higher Education; 1983.
7. Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology 1975;60(2):283-284. [doi: 10.1037/h0076540]
8. McLaughlin GH. SMOG grading: A new readability formula. Journal of Reading 1969;12(8):639-646.
9. Senter R, Smith E. Automated readability index. Aerospace Medical Research Laboratories 1967:1-14.
10. Liu Y, Ji M, Lin SS, Zhao M, Lyv Z. Combining Readability Formulas and Machine Learning for Reader-oriented Evaluation of Online Health Resources. IEEE Access 2021;9:67610-67619. [doi: 10.1109/access.2021.3077073]
11. Borst A, Gaudinat A, Grabar N, Boyer C. Lexically-based distinction of readability levels of health documents. Acta Informatica Medica 2008;16(2):72-75.
12. Zheng J, Yu H. Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study. J Med Internet Res 2017 Mar 02;19(3):e59 [FREE Full text] [doi: 10.2196/jmir.6962] [Medline: 28254738]
13. Horner SD, Surratt D, Juliusson S. Improving Readability of Patient Education Materials. Journal of Community Health Nursing 2000 Mar;17(1):15-23. [doi: 10.1207/s15327655jchn1701_02]
14. Vahabi M, Ferris L. Improving written patient education materials: a review of the evidence. Health Education Journal 2016 Jul 27;54(1):99-106. [doi: 10.1177/0017869969505400110]
15. Taylor HE, Bramley DEP. An Analysis of the Readability of Patient Information and Consent forms used in Research Studies in Anaesthesia in Australia and New Zealand. Anaesth Intensive Care 2012 Nov 01;40(6):995-998. [doi: 10.1177/0310057x1204000610]
16. Mumford M. A descriptive study of the readability of patient information leaflets designed by nurses. J Adv Nurs 1997 Nov;26(5):985-991. [doi: 10.1046/j.1365-2648.1997.00455.x] [Medline: 9372404]
17. Karačić J, Dondio P, Buljan I, Hren D, Marušić A. Languages for different health information readers: multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats. BMC Med Res Methodol 2019 Apr 05;19(1):75 [FREE Full text] [doi: 10.1186/s12874-019-0716-x] [Medline: 30953453]
18. Kim W, Kim I, Baltimore K, Imtiaz AS, Bhattacharya BS, Lin L. Simple contents and good readability: Improving health literacy for LEP populations. Int J Med Inform 2020 Sep;141:104230. [doi: 10.1016/j.ijmedinf.2020.104230] [Medline: 32688291]
19. Wilson M. Readability and patient education materials used for low-income populations. Clinical Nurse Specialist 2009;23(1):33-40. [doi: 10.1097/01.nur.0000343079.50214.31]
20. Balyan R, Crossley SA, Brown W, Karter AJ, McNamara DS, Liu JY, et al. Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. PLoS One 2019 Feb 22;14(2):e0212488 [FREE Full text] [doi: 10.1371/journal.pone.0212488] [Medline: 30794616]
21. Andersen EB. Introduction to the Statistical Analysis of Categorical Data. Berlin/Heidelberg, Germany: Springer Science & Business Media; 2012.
22. Lison P. An introduction to machine learning. University of Oslo. 2012. URL: http://home.nr.no/~plison/pdfs/talks/machinelearning.pdf [accessed 2021-10-10]
23. Australian Government Department of Health. URL: https://www.health.gov.au/ [accessed 2021-03-20]
24. Government of Western Australia Department of Health. URL: https://ww2.health.wa.gov.au/ [accessed 2021-03-20]
25. NSW Health. URL: https://www.health.nsw.gov.au/ [accessed 2021-03-20]
26. Ayub M, El-Alfy ESM. Impact of Normalization on BiLSTM Based Models for Energy Disaggregation. 2020 Presented at: 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI); October 26-27, 2020; Sakheer, Bahrain. [doi: 10.1109/icdabi51230.2020.9325593]
27. Singh D, Singh B. Investigating the impact of data normalization on classification performance. Applied Soft Computing 2020 Dec;97:105524. [doi: 10.1016/j.asoc.2019.105524]
28. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform. 2014. URL: https://www.frontiersin.org/articles/10.3389/fninf.2014.00014/full [accessed 2021-03-20]
29. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H. xgboost: Extreme Gradient Boosting. R Project. URL: https://cran.r-project.org/web/packages/xgboost/index.html [accessed 2021-10-10]
30. UCREL Semantic Analysis System (USAS). UCREL. URL: http://ucrel.lancs.ac.uk/usas/ [accessed 2021-03-20]
31. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46(1):389-422 [FREE Full text]

32. Siddharthan A. Syntactic Simplification and Text Cohesion. Res Lang Comput 2006 Mar 28;4(1):77-109 [FREE Full text] [doi: 10.1007/s11168-006-9011-1]
33. Siddharthan A. A survey of research on text simplification. ITL 2015 Jan 23;165(2):259-298 [FREE Full text] [doi: 10.1075/itl.165.2.06sid]
34. Poornima C, Dhanalakshmi V, Anand Kumar M, Soman KP. Rule based Sentence Simplification for English to Tamil Machine Translation System. IJCA 2011 Jul 31;25(8):38-42 [FREE Full text] [doi: 10.5120/3050-4147]
35. Arfé B, Mason L, Fajardo I. Simplifying informational text structure for struggling readers. Read Writ 2017 Oct 24;31(9):2191-2210 [FREE Full text] [doi: 10.1007/s11145-017-9785-6]
36. Crossley S, Allen D, McNamara DS. Text simplification and comprehensible input: A case for an intuitive approach. Language Teaching Research 2011 Dec 12;16(1):89-108 [FREE Full text] [doi: 10.1177/1362168811423456]
37. Britton B, Glynn SM, Meyer BJ, Penland MJ. Effects of text structure on use of cognitive capacity during reading. Journal of Educational Psychology 1982;74(1):51-61 [FREE Full text] [doi: 10.1037/0022-0663.74.1.51]
38. Bordag D, Kirschenbaum A, Tschirner E, Opitz A. Incidental acquisition of new words during reading in L2: Inference of meaning and its integration in the L2 mental lexicon. Bilingualism 2014 Jun 10;18(3):372-390 [FREE Full text] [doi: 10.1017/s1366728914000078]
39. Gooding S, Kochmar E, Yimam SM, Biemann C. Word Complexity is in the Eye of the Beholder. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2021; Online p. 4439-4449. [doi: 10.18653/v1/2021.naacl-main.351]
40. Reed R, Marks RJ. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks. Cambridge, MA, USA: MIT Press; 1999.

Abbreviations

ALL: all initial
AS: automatically selected
AUC: area under the curve
LE: linguistically enhanced
LR: logistic regression
RBF: radial basis function
RF: random forest
RFE: recursive feature elimination
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by G Eysenbach; submitted 07.04.21; peer-reviewed by K Allen; comments to author 20.05.21; revised version received 02.06.21; accepted 18.09.21; published 26.10.21.

Please cite as:
Xie W, Ji C, Hao T, Chow CY
Predicting the Easiness and Complexity of English Health Materials for International Tertiary Students With Linguistically Enhanced Machine Learning Algorithms: Development and Validation Study
JMIR Med Inform 2021;9(10):e25110
URL: https://medinform.jmir.org/2021/10/e25110
doi:10.2196/25110
PMID:34698644

Corrigenda and Addenda

Correction: Evaluation of Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study

Christina Schüttler[1], MSc; Hans-Ulrich Prokosch[1], PhD; Martin Sedlmayr[2], PhD; Brita Sedlmayr[2], PhD

[1]Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[2]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

Corresponding Author:
Christina Schüttler, MSc
Chair of Medical Informatics
Friedrich-Alexander-Universität Erlangen-Nürnberg
Wetterkreuz 15
Erlangen, 91058
Germany
Phone: 49 91318567789
Email: christina.schuettler@fau.de

Related Article:

Correction of: https://medinform.jmir.org/2021/7/e25531

(JMIR Med Inform 2021;9(10):e33105) doi:10.2196/33105

In "Evaluation of Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study" (JMIR Med Inform 2021;9(7):e25531) the authors noted one error.

In the originally published manuscript, a funding code was not included in the "Acknowledgments" section. The following sentence has now been added to this section:

> *This work is additionally supported by the German Federal Ministry of Education and Research under the funding code 01EY1701.*

The complete "Acknowledgments" section originally read as follows:

> *The authors would like to thank all participating MIRACUM locations—Dresden, Erlangen, Frankfurt am Main, Freiburg, Gießen, Greifswald, Magdeburg, and Mannheim. The authors would like to specially thank all scientists and researchers who participated in this study and provided a valuable insight into the usability of the different query builders through their loud thoughts and questionnaire evaluations. The authors would also like to thank Stefanie Schild (Erlangen), Renate Häuslschmid (Freiburg), and Preetha Moorthy (Mannheim) for their support in pretesting the tasks and questionnaires. This study was conducted as part of MIRACUM. MIRACUM is funded by the German Federal Ministry of Education and Research within the Medical Informatics Funding Scheme under the funding codes 01ZZ1801A and 01ZZ1801L.*
>
> *The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (CS).*

This has been corrected to:

> *The authors would like to thank all participating MIRACUM locations—Dresden, Erlangen, Frankfurt am Main, Freiburg, Gießen, Greifswald, Magdeburg, and Mannheim. The authors would like to specially thank all scientists and researchers who participated in this study and provided a valuable insight into the usability of the different query builders through their loud thoughts and questionnaire evaluations. The authors would also like to thank Stefanie Schild (Erlangen), Renate Häuslschmid (Freiburg), and Preetha Moorthy (Mannheim) for their support in pretesting the tasks and questionnaires. This study was conducted as part of MIRACUM. MIRACUM is funded by the German Federal Ministry of Education and Research within the Medical Informatics Funding Scheme under the funding codes 01ZZ1801A and 01ZZ1801L. This work is additionally supported by the German Federal Ministry of Education and Research under the funding code 01EY1701.*
>
> *The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol.*

hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (CS).

The correction will appear in the online version of the paper on the JMIR Publications website on October 8, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 24.08.21; this is a non–peer-reviewed article;accepted 06.09.21; published 08.10.21.

Please cite as:
Schüttler C, Prokosch HU, Sedlmayr M, Sedlmayr B
Correction: Evaluation of Three Feasibility Tools for Identifying Patient Data and Biospecimen Availability: Comparative Usability Study
JMIR Med Inform 2021;9(10):e33105
URL: https://medinform.jmir.org/2021/10/e33105
doi:10.2196/33105
PMID:34623958

Editorial

Health Natural Language Processing: Methodology Development and Applications

Tianyong Hao[1], PhD; Zhengxing Huang[2], PhD; Likeng Liang[1], PhD; Heng Weng[3], PhD; Buzhou Tang[4], PhD

[1]School of Computer Science, South China Normal University, Guangzhou, China

[2]College of Computer Science and Technology, Zhejiang University, Guangzhou, China

[3]The Second Affiliated Hospital, Guangzhou University of Chinese Medicine, Guangzhou, China

[4]School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, Shenzhen, China

Corresponding Author:
Buzhou Tang, PhD
School of Computer Science and Technology
Harbin Institute of Technology Shenzhen
L1819, Harbin Institute of Technology Campus, Xili University Town, Nanshan
Shenzhen, 518055
China
Phone: 86 0755 26033182
Email: tangbuzhou@gmail.com

Abstract

With the rapid growth of information technology, the necessity for processing substantial amounts of health data using advanced information technologies is increasing. A large amount of valuable data exists in natural text such as diagnosis text, discharge summaries, online health discussions, and eligibility criteria of clinical trials. Health natural language processing, as an interdisciplinary field of natural language processing and health care, plays a substantial role in a wide scope of both methodology development and applications. This editorial shares the most recent methodology innovations of health natural language processing and applications in the medical domain published in this JMIR Medical Informatics special theme issue entitled "Health Natural Language Processing: Methodology Development and Applications".

(JMIR Med Inform 2021;9(10):e23898) doi:10.2196/23898

KEYWORDS

health care; unstructured text; natural language processing; methodology; application

Introduction

Text data in an unstructured format widely exists in the medical domain, such as diagnosis records, operation records, discharge summaries, eligibility criteria of clinical trials, social media comments, online health discussions, and medical publications. Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) language texts. NLP aims to provide computer programs with the ability to process and understand unstructured text data. In the health arena, NLP techniques have been shown to be useful in dealing with information overload in the health and medical domain (eg, aggregation and summarization of patient notes, treatment analysis, information extraction and retrieval from massive discharge summaries, and semantic understanding of patient queries) [1]. NLP may also be applied for assisting medical decision-making by automatically analyzing the commonalities and differences of a large amount of text data and recommending appropriate actions on behalf of domain experts [2].

Health NLP, as an interdisciplinary field of NLP and health care, focuses on the methodology development of NLP and its applications in health care. It facilitates the analysis of the commonalities and differences of large amounts of text data and recommends appropriate actions on behalf of domain experts to assist medical decision-making. In general, it plays an essential role in processing various types of health text data and supports health applications to improve health care efficiency and efficacy.

With the increasing attention on this research field, there are more and more developments related to health NLP. Velupillai et al [3] shared the recent advances of health NLP in support of semantic analysis, covering the development of efficient methods for health corpus annotation/deidentification and the leverage of NLP for clinical utility including NLP infrastructure

for a clinical use case. Kalyan and Sangeetha [4] investigated the embeddings in health NLP for text representation in deep learning–based NLP tasks in clinical domains. The National NLP Clinical Challenges/Open Health Natural Language Processing (OHNLP) Competition [5] is held for family history extraction from synthetic clinical narratives using NLP. A number of health NLP tools and systems were also developed. For example, OHNLP released a catalog of clinical NLP software and provides interfaces to simplify the interaction of NLP systems [6]. Typically, Apache cTAKES [7], as an NLP system for extraction of information from electronic medical record clinical free text, aimed to integrate best-of-breed annotators, providing a world-class NLP system for accessing clinical information within the free text.

Methods

Health NLP covers a wide scope of methodology research including topics about methodology research such as NLP models for medical or social web data (eg, literature, EHRs, clinical trials, and social media about health care) processing; health information retrieval and extraction; NLP-assisted health information aggregation, abstraction, and summarization; machine learning–based text mining methods for health care; health care knowledge representation and reasoning; heath text corpora construction and annotation; and medical ontology and health knowledge graphs construction.

With respect to the application of NLP methods to the health care domain, health NLP contains NLP techniques for medicine personalization; question answering technologies for health applications; novel tools for medical, clinical, or social web data interpretation and visualization; innovative NLP systems for mobile environments for health care applications; NLP for clinical decision support and informatics; and applications of advanced NLP methods in clinical practice.

We reviewed 10 published articles in the JMIR Medical Informatics theme issue Health Natural Language Processing: Methodology Development and Applications to share the recent developments of the studies, from methodology research to applications.

Results

Medical Information Extraction

Medical information extraction is a key technology that supports the development of medical informatics. Zhang et al [8] developed a new Chinese electronic medical record (EMR) data set with six types of entities and proposed a multilevel representation learning model based on Bidirectional Encoder Representation from Transformers (BERT) for Chinese medical entity recognition. The experiments on the Chinese EMR data set and China Conference on Knowledge Graph and Semantic Computing 2018 benchmark data set showed that the proposed method outperformed state-of-the-art methods. Automatic relation extraction between chemicals and diseases plays an important role in biomedical text mining. Wang et al [9] proposed an end-to-end neural network based on a graph convolutional network and multi-head attention. To improve the performance, a document-level dependency graph was constructed to capture dependency syntactic information across sentences. The graph was applied to capture the feature representation of a document-level dependency graph, while the multi-head attention mechanism was used to learn relative context features from different semantic subspaces. The experiment results showed that the method achieved the best F-score, which was superior to state-of-the-art methods. The graph convolutional network model was effectively used for dependency information across sentences to improve the performance of intersentence chemical-disease extraction. Targeted at extracting the interactions between chemicals and proteins from the biomedical literature, Wang et al [10] proposed effectively encoding syntactic information from long text. The method leveraged graph convolutional networks to capture sequential information and long-range syntactic relations between words by using the dependency structure of input sentences. The evaluation of the ChemProt corpus showed that the model achieved an F-score of 65.17%, which was 1.07% higher than that of the state-of-the-art system. The study indicated that the graph neural network–based model could better capture the semantic and syntactic information of the biomedical literature sentence. Temporal information frequently exists in the representation of the disease progress, prescription, medication, surgery progress, or discharge summary in narrative clinical text. The extraction and normalization of temporal expressions can positively boost the analysis and understanding of narrative clinical texts to promote clinical research and practice. Pan et al [11] proposed a rule-based and pattern learning–based model for extracting and normalizing temporal expressions from Chinese narrative clinical text. The model consisted of three stages: extraction, classification, and normalization. Based on a set of narrative clinical texts in Chinese containing 1459 discharge summaries of a domestic Grade-A Class 3 hospital, the performance of the model achieved the performance compared with baseline methods. The research of medical information extraction still has the challenges of insufficient training data size, complex domain terminology, a large proportion of noise data, and significant inconsistency among various data types.

Heath Knowledge Graph and Its Applications

Targeted at knowledge graph embedding for semantic representation of entities and relations, the challenge of how to learn probability values of triplets into representation vectors was addressed. Li et al [12] constructed a mapping function between score value and probability, and introduced probability-based loss of triplets into original margin-based loss function. Compared with state-of-the-art TransX algorithms, the proposed model performed better in all evaluation indicators. Checking whether the medication is clinically reasonable with respect to the diagnosis is the key to fraud, waste, and abuse detection, which is a significant yet challenging problem in the health insurance industry. Sun et al [13] built an automatic method to identify the clinically suspected claims for fraud, waste, and abuse detection by using a medical knowledge graph. A deep learning–based method was applied to extract the entities and relationships from knowledge sources, and a multilevel similarity matching method was developed for entity linking.

From 185,796 drug labels from the China Food and Drug Administration, a medical knowledge graph containing 1,616,549 nodes and 5,963,444 edges was constructed for identifying fraud, waste, and abuse suspects. The research of health knowledge graphs still has the challenges of complex text representation, low extract performance, and limited knowledge graph size.

NLP Methods for Health Text Mining

Traditional Chinese medicine (TCM) has been shown to be an efficient mode to manage advanced lung cancer, and accurate syndrome differentiation is crucial to treatment. Liu et al [14] established five deep learning–based TCM diagnostic models to imitate lung cancer syndrome differentiation. The models used unstructured medical records as inputs to capitalize on data collected for practical TCM treatment cases by lung cancer experts. The experiment result showed the F1-score of the recurrent convolutional neural network model improved over models without data augment. The text-hierarchical attention network model achieved the highest F1-score. Medical records could be used more productively by constructing end-to-end models to facilitate lung cancer. The classification of clinical trial eligibility criteria texts is a fundamental and critical step in clinical target population recruitment. Zeng et al [15] proposed an ensemble learning method that integrates the current cutting-edge deep learning models BERT, Enhanced Language Representation with Informative Entities, XLNet, and RoBERT. Through a model ensemble in two layers, the study trained a model and compared it with a list of baseline deep learning models on a publicly available standard data set. The results demonstrated that the proposed ensemble learning method outperformed a list of baseline methods. The research of NLP methods still heavily relies on the advancement of machine learning models.

Advanced Applications

Deidentification of clinical records, as an application, is a critical step in the use of electronic health records for academic research. Zhao et al [16] investigated the usefulness of rule-based learners in a hybrid deidentification system. A data-driven rule learner named transformation-based error-driven learning was integrated into a hybrid system. Based on the widely used Informatics for Integrating Biology and the Bedside deidentification data set, the learner could offer high performance with generated rules. After integrating the learner into an ensemble framework, the performance achieved the best among the community. The rule-based method thus could offer an effective contribution to the current ensemble learning approach for the deidentification of clinical records as a typical application in medical informatics. An artificial intelligence–based assistive diagnostic system is designed to diagnose multiple types of diseases that are common in TCM based on patients' electronic health record notes. Zhang et al [17] developed a method to simultaneously diagnose the disease and produce a list of corresponding syndromes. NLP techniques using a recurrent neural network model were applied to process unstructured electronic health record notes to extract clinical information such as signs and symptoms that were represented by named entities. A total of 22,984 electronic health records from Guanganmen Hospital of the China Academy of Chinese Medical Sciences were collected and applied to the diagnostic system. From the evaluation, 187 commonly known TCM diseases could be diagnosed, and a wider range of TCM disease types was expected to be diagnosed. The applications of NLP methods tend to be more and more widespread in the health care domain. However, the challenges, including the security of data, the actual needs from clinicians, the validation of results, and user convenience, still need to be solved in the future.

Conclusion

Health NLP draws more and more attention for its essential role in a wide scope of both methodology development and applications. This editorial shares the most recent methodology research of health NLP and its applications in health care by reviewing 10 newly published articles on the JMIR Medical Informatics theme issue Health Natural Language Processing: Methodology Development and Applications. The research indicates recent focuses on medical information extraction (entity, relation, temporal, and interaction extraction), knowledge graph construction and use, methods for clinical decision support and informatics, and NLP systems for health care applications in practice.

Acknowledgments

TH is supported by the National Natural Science Foundation of China (61772146) and Guangzhou Science Technology and Innovation Commission (201803010063). ZH is supported by the National Natural Science Foundation of China (61672450). HW is supported by the National Natural Science Foundation of China (61871141). BT is supported by the following grants: National Natural Science Foundations of China (U1813215 and 61876052); Special Foundation for Technology Research Program of Guangdong Province (2015B010131010); National Natural Science Foundation of Guangdong, China (2019A1515011158); Guangdong Province Covid-19 Pandemic Control Research Fund (2020KZDZX1222); Strategic Emerging Industry Development Special Fund of Shenzhen (JCYJ20180306172232154); and Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052).

Conflicts of Interest

None declared.

References

1. Chen X, Xie H, Wang F, Liu Z, Xu J, Hao T. A bibliometric analysis of natural language processing in medical research. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):14 [FREE Full text] [doi: 10.1186/s12911-018-0594-x] [Medline: 29589569]
2. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. J Biomed Inform 2014 Dec;52:112-120 [FREE Full text] [doi: 10.1016/j.jbi.2014.01.009] [Medline: 24496068]
3. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015 Aug 13;10(1):183-193 [FREE Full text] [doi: 10.15265/IY-2015-009] [Medline: 26293867]
4. Kalyan KS, Sangeetha S. SECNLP: a survey of embeddings in clinical natural language processing. J Biomed Inform 2020 Jan;101:103323 [FREE Full text] [doi: 10.1016/j.jbi.2019.103323] [Medline: 31711972]
5. Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, et al. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR Med Inform 2021 Jan 27;9(1):e24008 [FREE Full text] [doi: 10.2196/24008] [Medline: 33502329]
6. Masanz J, Pakhomov SV, Xu H, Wu ST, Chute CG, Liu H. Open source clinical NLP - more than any single system. AMIA Jt Summits Transl Sci Proc 2014;2014:76-82 [FREE Full text] [Medline: 25954581]
7. Apache cTAKES. URL: https://ctakes.apache.org/ [accessed 2021-02-01]
8. Zhang Z, Zhu L, Yu P. Multi-level representation learning for Chinese medical entity recognition: model development and validation. JMIR Med Inform 2020 May 04;8(5):e17637 [FREE Full text] [doi: 10.2196/17637] [Medline: 32364514]
9. Wang J, Chen X, Zhang Y, Zhang Y, Wen J, Lin H, et al. Document-level biomedical relation extraction using graph convolutional network and multihead attention: algorithm development and validation. JMIR Med Inform 2020 Jul 31;8(7):e17638 [FREE Full text] [doi: 10.2196/17638] [Medline: 32459636]
10. Wang E, Wang F, Yang Z, Wang L, Zhang Y, Lin H, et al. A graph convolutional network-based method for chemical-protein interaction extraction: algorithm development. JMIR Med Inform 2020 May 19;8(5):e17643 [FREE Full text] [doi: 10.2196/17643] [Medline: 32348257]
11. Pan X, Chen B, Weng H, Gong Y, Qu Y. Temporal expression classification and normalization from Chinese narrative clinical texts: pattern learning approach. JMIR Med Inform 2020 Jul 27;8(7):e17652 [FREE Full text] [doi: 10.2196/17652] [Medline: 32716307]
12. Li L, Wang P, Wang Y, Wang S, Yan J, Jiang J, et al. A method to learn embedding of a probabilistic medical knowledge graph: algorithm development. JMIR Med Inform 2020 May 21;8(5):e17645 [FREE Full text] [doi: 10.2196/17645] [Medline: 32436854]
13. Sun H, Xiao J, Zhu W, He Y, Zhang S, Xu X, et al. Medical knowledge graph to enhance fraud, waste, and abuse detection on claim data: model development and performance evaluation. JMIR Med Inform 2020 Jul 23;8(7):e17653 [FREE Full text] [doi: 10.2196/17653] [Medline: 32706714]
14. Liu Z, He H, Yan S, Wang Y, Yang T, Li GZ. End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: model development and validation. JMIR Med Inform 2020 Jun 16;8(6):e17821 [FREE Full text] [doi: 10.2196/17821] [Medline: 32543445]
15. Zeng K, Pan Z, Xu Y, Qu Y. An ensemble learning strategy for eligibility criteria text classification for clinical trial recruitment: algorithm development and validation. JMIR Med Inform 2020 Jul 01;8(7):e17832 [FREE Full text] [doi: 10.2196/17832] [Medline: 32609092]
16. Zhao Z, Yang M, Tang B, Zhao T. Re-examination of rule-based methods in deidentification of electronic health records: algorithm development and validation. JMIR Med Inform 2020 Apr 30;8(4):e17622 [FREE Full text] [doi: 10.2196/17622] [Medline: 32352384]
17. Zhang H, Ni W, Li J, Zhang J. Artificial intelligence-based traditional Chinese medicine assistive diagnostic system: validation study. JMIR Med Inform 2020 Jun 15;8(6):e17608 [FREE Full text] [doi: 10.2196/17608] [Medline: 32538797]

Abbreviations

BERT: Bidirectional Encoder Representation from Transformers
NLP: Natural Language Processing
OHNLP: Open Health Natural Language Processing
TCM: Traditional Chinese Medicine

Edited by C Lovis; submitted 27.08.20; peer-reviewed by M Cai, J Ainsworth; comments to author 11.10.20; revised version received 28.02.21; accepted 27.04.21; published 21.10.21.

Please cite as:
Hao T, Huang Z, Liang L, Weng H, Tang B
Health Natural Language Processing: Methodology Development and Applications
JMIR Med Inform 2021;9(10):e23898
URL: https://medinform.jmir.org/2021/10/e23898
doi:10.2196/23898
PMID:34673533

Original Paper

Harnessing the Electronic Health Record and Computerized Provider Order Entry Data for Resource Management During the COVID-19 Pandemic: Development of a Decision Tree

Hung S Luu[1], PharmD, MD; Laura M Filkins[1], PhD; Jason Y Park[1], MD, PhD; Dinesh Rakheja[1], MD; Jefferson Tweed[2], MS; Christopher Menzies[2], MD; Vincent J Wang[3], MD, MHA; Vineeta Mittal[4], MD, MBA; Christoph U Lehmann[5,6,7,8], MD; Michael E Sebert[9], MD

[1]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, United States
[2]Department of Advanced Analytics and Informatics, Children's Health, Dallas, TX, United States
[3]Division of Pediatric Emergency Medicine, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, United States
[4]Division of Pediatric Hospital Medicine, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, United States
[5]Clinical Informatics Center, University of Texas Southwestern Medical Center, Dallas, TX, United States
[6]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, United States
[7]Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, United States
[8]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, United States
[9]Division of Infectious Diseases, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, United States

Corresponding Author:
Hung S Luu, PharmD, MD
Department of Pathology
University of Texas Southwestern Medical Center
1935 Medical District Drive
Dallas, TX, 75235
United States
Phone: 1 2144562168
Fax: 1 2144564713
Email: hung.luu@childrens.com

Abstract

Background: The COVID-19 pandemic has resulted in shortages of diagnostic tests, personal protective equipment, hospital beds, and other critical resources.

Objective: We sought to improve the management of scarce resources by leveraging electronic health record (EHR) functionality, computerized provider order entry, clinical decision support (CDS), and data analytics.

Methods: Due to the complex eligibility criteria for COVID-19 tests and the EHR implementation–related challenges of ordering these tests, care providers have faced obstacles in selecting the appropriate test modality. As test choice is dependent upon specific patient criteria, we built a decision tree within the EHR to automate the test selection process by using a branching series of questions that linked clinical criteria to the appropriate SARS-CoV-2 test and triggered an EHR flag for patients who met our institutional persons under investigation criteria.

Results: The percentage of tests that had to be canceled and reordered due to errors in selecting the correct testing modality was 3.8% (23/608) before CDS implementation and 1% (262/26,643) after CDS implementation (P<.001). Patients for whom multiple tests were ordered during a 24-hour period accounted for 0.8% (5/608) and 0.3% (76/26,643) of pre- and post-CDS implementation orders, respectively (P=.03). Nasopharyngeal molecular assay results were positive in 3.4% (826/24,170) of patients who were classified as asymptomatic and 10.9% (1421/13,074) of symptomatic patients (P<.001). Positive tests were more frequent among asymptomatic patients with a history of exposure to COVID-19 (36/283, 12.7%) than among asymptomatic patients without such a history (790/23,887, 3.3%; P<.001).

Conclusions: The leveraging of EHRs and our CDS algorithm resulted in a decreased incidence of order entry errors and the appropriate flagging of persons under investigation. These interventions optimized reagent and personal protective equipment

usage. Data regarding symptoms and COVID-19 exposure status that were collected by using the decision tree correlated with the likelihood of positive test results, suggesting that clinicians appropriately used the questions in the decision tree algorithm.

(JMIR Med Inform 2021;9(10):e32303) doi:10.2196/32303

KEYWORDS

COVID-19; computerized provider order entry; electronic health record; resource utilization; personal protective equipment; SARS-CoV-2 testing; clinical decision support

Introduction

COVID-19 is caused by SARS-CoV-2 and has quickly emerged as a global pandemic since its initial description in December 2019 [1]. The increased testing and isolation of patients with COVID-19 are important means of limiting the spread of infection. Many laboratories in the United States have expanded their testing capabilities rapidly [2]. As a result, the overall testing capacity in the United States is substantially larger than what the Centers for Disease Control and Prevention and state health agencies were able provide at the start of the pandemic. Testing shortages however persisted throughout 2020 and, to a lesser extent, into 2021 due to inadequate supplies of collection swabs, viral transport media, RNA extraction regents, and other reagents and consumables [3,4]. Institutions have had to prioritize testing by taking into account the severity of illnesses, the rapidness of results, bed availability, and staffing needs [4].

Electronic health records (EHRs) and computerized provider order entry (CPOE) systems offer the potential to reduce the number of medical errors and improve care quality by facilitating communication, providing access to information, monitoring patients, providing decision support, and enhancing clinicians' situational awareness [5-7]. However, EHRs can also inadvertently result in clinicians introducing new errors, overlooking existing orders, and duplicating work [8-10]. Apart from the need to reduce costs, preventing the duplicate testing of patients for COVID-19 is essential for conserving existing testing supplies and maximizing the number of patients that can be tested.

Although the availability of testing is important, so is the timely dissemination of test results to care providers to optimally allocate valuable hospital resources, such as limited supplies of personal protective equipment (PPE), effectively [4]. Testing capacities have increased since the early days of the pandemic, but the proliferation of different testing platforms and methodologies has led to variations in test turnaround times and assay sensitivity. Commercial vendors have produced high-throughput, cartridge-based instruments that promise shorter testing turnaround times; however, the demand for these instruments currently exceeds the amount of available supplies [4].

To meet the testing needs of our patient population despite equipment shortages, institutions such as our pediatric health care system had to assemble a variety of COVID-19 testing modalities with varying performance characteristics. Matching testing modalities to the appropriate clinical scenario was a challenge. Some institutions developed decision-making algorithms to stratify their patient population into risk groupings [11]. Herein, we describe and evaluate the CPOE clinical decision support (CDS) tools that were developed to optimize the ordering of COVID-19 tests; the EHR functionalities that were leveraged to manage persons under investigation (PUIs); and the data analysis tools that were essential for monitoring changing variables, such as ordering patterns and available reagent supplies.

Methods

Setting and Institutional Approach to Managing the COVID-19 Pandemic

Our academically affiliated pediatric health care system in North Texas consists of 3 acute care hospitals that are licensed for a total of 601 beds and 24 ambulatory specialty care centers. Together, these facilities care for more than 227,000 unique patients per year and have provided services, including more than 19,600 surgeries and 107,800 emergency department visits [12]. Our health system's efforts in preparing for patients with COVID-19 began early in 2020 and included the activation of the Hospital Incident Command Structure on March 5. A sick isolation unit was opened on March 23 for the management of patients who did not require critical care and were either suspected of SARS-CoV-2 infection—designated as PUIs—or confirmed to be infected. The first positive SARS-CoV-2 test result for a patient in our system was received later that month (March 31). With the activation of the Hospital Incident Command Structure, we recognized that the pandemic would require an organized, sustainable, and adaptable approach to caring for children with COVID-19 while minimizing staff exposure and optimizing the use of PPE and testing reagents and supplies. In this study, we describe and evaluate tools that were developed within the EHR and were vital components of this approach.

As COVID-19 spread across the world and within the United States, the epidemiology of the disease morphed over time. First, cases were seen predominately among patients who had been exposed to the disease during recent travel. Afterward, the disease began to spread within communities, but most new infections were still identified among individuals who had contact with a limited number of confirmed local cases. Finally, widespread community transmission developed, and many cases could no longer be reliably related to a known exposure or travel history [13-15]. In early 2020, the criteria recommended by the Centers for Disease Control and Prevention for identifying a person as a PUI changed several times [16,17]. Reflecting the changing disease epidemiology, these PUI definitions, which had initially focused on symptomatic individuals with a history of travel to Wuhan, China, or a history of contact with a

laboratory-confirmed case of COVID-19, were later expanded by the addition of criteria related to travel from mainland China, travel from affected geographic areas within the United States, and, finally, even individuals with no known exposure risk factors [16]. Following the initial pandemic period, during which SARS-CoV-2 testing was available at our institution only through public health laboratories, the options for testing increased first thanks to offerings from commercial reference laboratories and then due to the launch of an internal, laboratory-developed test with a turnaround time of approximately 24 hours. Later, our laboratory implemented commercial rapid testing platforms that offered further improvements in turnaround times for a limited number of specimens depending on the availability of the required kits (Table 1).

Table 1. The SARS-CoV-2 assays implemented.

Assay characteristic	Modified CDC[a] SARS-CoV-2 Assay (laboratory-developed test)	Biofire Respiratory Panel 2.1 (bioMérieux SA)	Xpert Xpress SARS-CoV-2 (Cepheid)	SARS[b] Antigen FIA[c] (Quidel Corporation)	Alinity m SARS-COV-2 Assay (Abbott Laboratories)[d]	Cobas SARS-CoV-2 (Roche Holding AG)[e]
Analyte	RNA	RNA	RNA	Antigen	RNA	RNA
Sample collection	NP[f] swab in UTM[g]	NP swab in UTM	NP swab in UTM	Anterior nares swab	NP swab in UTM	NP swab in UTM
SARS-CoV-2 target	Nucleocapsid gene	Membrane gene and surface gene	Envelope gene and nucleocapsid 2 gene	Nucleocapsid protein	Nucleocapsid gene and RdRp[h] gene	Envelope gene and RdRp gene
SARS-CoV-2 LoD[i]	260 copies/mL	160 copies/mL	250 copies/mL	113 TCID50[j]/mL	100 copies/mL	0.003 TCID50/mL
Other target(s)	None	21 additional viruses and bacteria	None	None	None	None
Instrument(s)	EMAG (extraction; bioMérieux SA) and ABI[k] 7500 (polymerase chain reaction; Thermo Fisher Scientific)	FilmArray Torch System (bioMérieux SA)	GeneXpert XVI (Cepheid)	Sofia 2 (Quidel Corporation)	Alinity m System (Abbott Laboratories)	Cobas 6800 (Roche Holding AG)
Maximum throughput[l]	150 samples/8-hour shift (extraction and polymerase chain reaction)	<1 hour/test/ instrument module	<1 hour/test/ instrument module	20 min/test/ instrument module	300 tests/8-hour shift	864 tests/8-hour shift
Time to results, mean (SD)[m]	0.79 (0.85) days	70 (17) min	77 (29) min	27 (5) min	0.53 (0.35) days	2.03 (1.56) days

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

[b]SARS: severe acute respiratory syndrome.

[c]FIA: fluorescent immunoassay.

[d]The assay was performed at reference lab 1.

[e]The assay was performed at reference lab 2.

[f]NP: nasopharyngeal.

[g]UTM: universal transport medium.

[h]RdRp: RNA-dependent RNA polymerase.

[i]LoD: limit of detection (the LoD shown is either the lowest reported [highest sensitivity] value on the package insert or the lowest value observed in the laboratory).

[j]TCID50: median tissue culture infectious dose.

[k]ABI: Applied Biosystems.

[l]Maximum throughput assumes sufficient reagents. Maximum throughput volumes were not achieved for most platforms due to limited reagent allocations.

[m]The time from specimen (primary orders) or order (add-on orders) receipt in the lab to result reporting. This includes transport to outside labs (send-out testing only), laboratory processing, sample preparation, instrument time, and result reporting.

New institutional policies and procedures, in response to the COVID-19 pandemic, were instituted in parallel with the changed understanding of the disease's epidemiology, the illness, and SARS-CoV-2 transmission. These changes included the adoption (on April 28, 2020) of universal SARS-CoV-2 testing for all patients who were admitted through the emergency department or directly to inpatient floors and the intensive care unit. At first, rapid testing was prioritized for patients with fevers or respiratory symptoms or those who had close contact with individuals with SARS-CoV-2 infection, while other patients were tested by using the laboratory-developed test. This strategy directed limited resources for rapid testing toward patients with

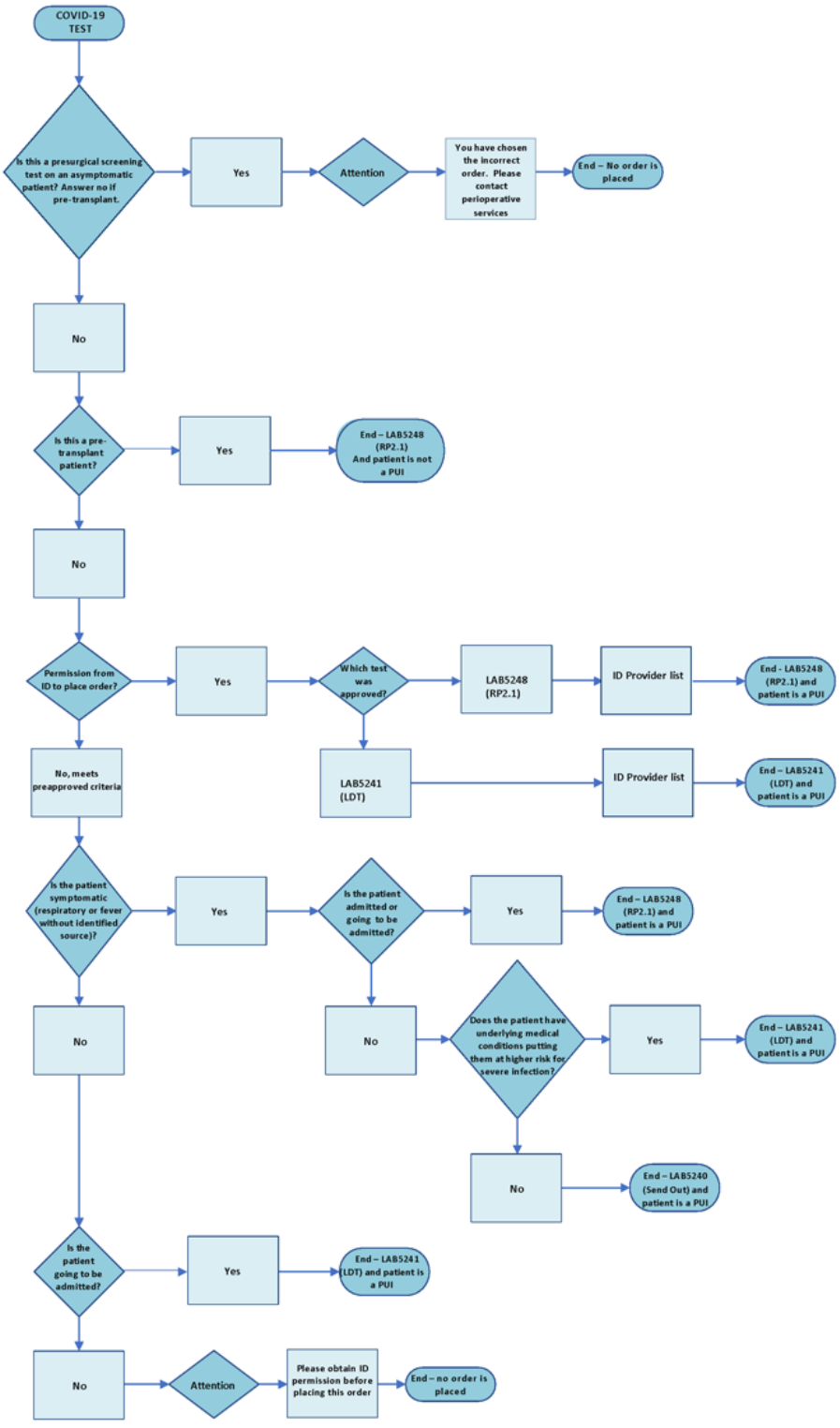
the highest likelihood of infection but resulted in a delay in identifying asymptomatic positive cases, which represent a considerable portion of SARS-CoV-2 infections in children. As rapid testing became increasingly available, such tests were deployed subsequently for all admitted patients.

To optimize the use of resources, such as negative pressure rooms and PPE, we developed a policy for aerosol-generating procedures (AGPs). The policy governed the performance of AGPs, including any preceding SARS-CoV-2 testing and PPE requirements, in a systematic manner that was driven by patients' symptoms, their COVID-19 status (if known), the prevalence of infection in the community, and the classification of AGPs into 2 risk tiers. SARS-CoV-2 testing was initially required in advance for all patients undergoing scheduled surgery, and the empiric use of PPE, including N95 respirators, was reserved for urgent or emergent cases when testing was not feasible. As community spread increased and access to rapid testing improved, the testing requirement was extended to any urgent surgical procedures for which sufficient time was available.

EHR Decision Tree for SARS-CoV-2 Test Order Placement

Given the scarcity of testing resources and growing demand during the early phase of the pandemic, formal criteria for SARS-CoV-2 testing were developed at our institution through consensus among physician and clinical laboratory leaders. Prior to the pandemic, our institution did not restrict the ordering of assays for non–SARS-CoV-2 respiratory viruses nor collect data on the reasons for ordering such tests systematically. Developing an ordering system that would be intuitive for clinicians to use and would capture data to guide the prioritization of orders and subsequent revisions to indications for ordering were therefore important priorities. However, the criteria for ordering specific COVID-19 tests were complex, and the metadata were frequently revised as new clinical scenarios were incorporated and new testing options became available. The implementation of the detailed ordering criteria in the EHR posed a challenge that increased with the number of available testing options. More importantly, the growing list of testing indications was a hard-to-navigate obstacle for care providers who needed to place orders. To ease the burden of ordering the correct test from a long list of choices, we built a decision tree within the EHR to automate the selection process based on answers that are provided to a branching set of hierarchical questions. This decision tree (Figure 1) was first implemented on April 28, 2020, and was subsequently updated and frequently modified during the early response to the pandemic.

Figure 1. Electronic health record decision tree for ordering SARS-CoV-2 tests. This flow diagram shows the branching set of hierarchical questions that resulted in the capture of data for test prioritization and symptom status identification. LDT: laboratory-developed test; PUI: person of interest; RP2.1: BioFire Respiratory Panel 2.1.



PUI Flagging in the EHR

In addition to linking clinical indications to the appropriate SARS-CoV-2 test, the ordering process required setting a flag in the EHR for any patient who met our institutional PUI criteria. The flag alerted health care personnel to a patient's PUI status and the need to use PPE beyond those for standard precautions, including N95 respirators or powered air purifying respirators, when caring for patients.

Testing for an infectious disease usually suggests a clinical index of suspicion that, in itself, may justify flagging patients in the EHR for the possibility of being infected with that disease. In the case of COVID-19 however, institutional policies required SARS-CoV-2 testing upon admission or before surgery for all

patients, even in the absence of symptoms or exposure, thereby rendering the presence of an ordered test functionally meaningless as an indicator of clinical suspicion. Although some patients may be asymptomatic carriers and thus could expose the workforce, the pretest probability of infection in such patients was not expected to be above that of the general population. The universal usage of N95 respirators for all health care encounters during the pandemic was neither recommended or feasible, given the limited supplies. Therefore, our institution decided that patients without compatible symptoms or recent exposure to SARS-CoV-2 would not be designated as PUIs, even when routine testing is required by institutional screening protocols. Consequently, in addition to guiding the selection of the correct SARS-CoV-2 test, the decision tree needed to assign the appropriate PUI status to each patient based on the indication for testing.

The introduction of additional testing modalities with decreased sensitivity compared to that of molecular nasopharyngeal sample testing modalities presented another challenge. Although positive results from these less sensitive assays were considered reliable, negative results were not and required confirmation with a more sensitive molecular test. Accordingly, the EHR rules for the clearance of PUI flags were constructed to require a negative molecular test from a nasopharyngeal sample, even if the flag had originally been triggered by an order for a less sensitive screening test.

A flagging system was also created for displaying results from SARS-CoV-2 tests that had been performed at outside facilities with interoperable EHRs. Such outside test results were either flagged as being reliable and approved by our system's laboratory as being equivalent to internal testing results (by the Happy Together EHR collaborative, which includes Children's Health, Parkland Hospital, and University of Texas Southwestern Medical Center) or otherwise flagged as results for which equivalence to internal testing results could not be established. Whether flagged patients were being seen in the emergency department or were directly admitted to the wards, the availability of this information allowed bedside physicians to avoid unnecessary SARS-CoV-2 testing, thereby minimizing the waste of limited testing resources.

EHR Tools and the Maintenance of PPE Supplies

Like many US health care institutions, early in the pandemic, we recognized the potential for a shortfall in the critical PPE supplies required for the care of patients with COVID-19, including N95 respirators. Providing appropriate protection to health care workers while minimizing PPE consumption made the accurate identification and flagging of PUIs essential. Our supply of N95 respirators reached a nadir in late March—less than 14 days' worth of stock on hand overall and less than 7 days' worth of supply for the scarcest respirator size—but subsequently recovered. Although multiple concurrent strategies, including UV reprocessing and the enhanced scrutiny of N95 respirator usage, also contributed to the successful management of this shortfall, the proper assignment of PUI statuses was a critical component in the struggle to reduce PPE use. Improvements in the national supply of N95 respirators have since reduced the acute importance of these considerations, but the strategies developed during the COVID-19 pandemic for managing limited PPE supplies will be beneficial approaches to dealing with future resource challenges.

Results

SARS-CoV-2 Test Ordering Metrics

The frequencies with which orders for SARS-CoV-2 tests needed to be revised due to user error or had to be repeated were used as measures for the impact of the CDS tools. The percentage of tests that were canceled and reordered due to errors in selecting the correct testing modality was 3.8% (23/608) prior to CDS implementation and 1% (262/26,643) after the implementation of CDS (Fisher exact test: $P<.001$). The percentages of patients for whom multiple tests were ordered during a 24-hour period were 0.8% (5/608) and 0.3% (76/26,643) prior to and after CDS implementation, respectively, as of October 31, 2020 (Fisher exact test: $P=.03$).

SARS-CoV-2 Infection Frequency

If the information captured by the decision tree regarding the assignment of SARS-CoV-2 test modalities and PUI statuses accurately reflected the risk of infection, it would be expected that the incidence of positive test results would vary accordingly. Patients were classified as symptomatic or asymptomatic via the decision tree based on the presence or absence of a fever without an identified source or the presence of respiratory symptoms. Consistent with our expectations, the observed frequency of positive nasopharyngeal molecular assays for asymptomatic patients (826/24,170, 3.4%; Table 2) was significantly lower (Fisher exact test: $P<.001$) than that frequency for symptomatic patients (1421/13,074, 10.9%). Likewise, the incidence of positive test results was higher among asymptomatic patients with a history of exposure to an individual with COVID-19 (36/283, 12.7%) than among asymptomatic patients without such an exposure history (790/23,887, 3.3%; Fisher exact test: $P<.001$).

Table 2. SARS-CoV-2 testing volumes and results by ordering indication.

Testing indication category[a]	Testing volume[b], N	Positive tests[b], n (%)
Asymptomatic patients[c]	24,170	826 (3.4)
Preprocedural screening[d]	12,864	428 (3.3)
Admission screening	10,625	329 (3.1)
Screening before behavioral health placement	398	33 (8.3)
Admission screening of asymptomatic patients with a history of close contact with an individual with COVID-19	283	36 (12.7)
Symptomatic patients	13,074	1421 (10.9)
Admission screening or hospitalized patients	5573	433 (7.8)
Preprocedural screening[d]	298	31 (10.4)
Outpatients with risk factors for severe illness	307	48 (15.6)
Lower respiratory tract disease without an alternative explanation[e]	30	3 (10)
Symptomatic patient with a history of close contact with an individual with COVID-19[e]	3	0 (0)
Symptomatic patient without other specified criteria	6863	906 (13.2)
Symptom status not specified	15,341	1146 (7.5)
Preprocedural screening[d]	6796	177 (2.6)
Unrestricted send-out testing	5330	791 (14.8)
Testing approved by the Division of Infectious Diseases	535	72 (13.5)
Patient screening after health care exposure	89	2 (2.2)
Unclassified testing	2591	104 (4)
Total testing	52,585	3393 (6.5)

[a]The testing indication categories listed summarize a larger number of actual indications displayed in the electronic health record, which were dynamically modified over the course of the pandemic.

[b]Testing data cover the period from March 13, 2020, through March 24, 2021.

[c]Patients without fevers and without respiratory symptoms were classified as asymptomatic.

[d]Includes testing before surgery and other qualifying aerosol-generating procedures.

[e]These criteria were used only briefly during the early phase of the pandemic, after which test eligibility was expanded to include symptomatic patients and tests did not need to consider these criteria.

Another group of asymptomatic patients for whom we observed a significantly increased incidence of positive SARS-CoV-2 test results included patients awaiting behavioral health placement (33/398, 8.3%; other asymptomatic patients without a history of COVID-19 exposure: 757/23,489, 3.2%; $P<.001$). The reason for this increased positivity rate is unclear, but some of these patients likely had a history of prior infection and were referred to our facilities for repeated testing before behavioral health placement to assess for viral clearance. Furthermore, the behavior patterns of these patients may have included decreased adherence to prevention measures such as mask wearing and social distancing, which placed them at an increased infection risk.

Testing for symptomatic patients when resources were the most limited was initially targeted toward those who (1) required hospitalization, (2) had comorbid conditions that increased their risk for developing a serious illness, (3) had a history of COVID-19 exposure, or (4) had a lower respiratory tract infection without another explanation. As the availability of test reagents improved, test eligibility was expanded more broadly to include symptomatic patients, and several of these more specific indications were retired. However, clinicians continued to use the decision tree to identify hospitalized patients and those with risk factors for severe illness to prioritize such patients for rapid testing. All symptomatic patients were designated as PUIs, even when the decision tree did not require more detailed information.

Symptom status was not captured for a subset of test orders (15,341/52,585, 29.2%). Many of these tests were assays that were either sent out to off-site laboratories for nonhospitalized patients or collected as screening tests several days in advance of a scheduled procedure. In the first case, symptomatic patients were instructed to isolate at home pending the result of the test. In the second case, presurgical screening results were generally available by the time patients returned for surgery. The empiric assignment of PUI statuses in the EHR at the time of testing was therefore not prioritized for these patients. Since September 2020 however, improvements in implementation resulted in the consistent capturing of symptom information for ≥80% of tested patients every month.

To manage rare or unanticipated circumstances, our testing algorithm allowed physicians in the Division of Infectious Diseases to authorize testing for patients who exhibited testing indications outside of those that were approved and implemented in the EHR. Once off-site testing became unrestricted, this approval option was used primarily for requests for locally performed tests that offered a shorter turnaround time or for patients who exhibited clinical indications that favored a specific testing platform. This approval route was needed only for 1% (535/52,585) of orders, indicating that the decision tree effectively managed a large majority of scenarios and prevented the approval activity from becoming an excessive burden on the physicians who were tasked with evaluating these nonstandard requests. The yield of positive results from such tests that were approved by infectious disease physicians was high (72/535, 13.5%), as was the frequency of positive results among unrestricted send-out tests (791/5330, 14.8%). These high rates of positive results suggest that clinicians were applying appropriate judgement to selecting patients for testing when considering these ordering options.

Discussion

Principal Findings

During the period following the implementation of CDS for SARS-CoV-2 test ordering, we documented improvements in the number of cancelled and reordered tests as well as decreases in the number of patients who underwent unnecessary duplicate testing. The goals of CPOE systems include submitting appropriate and efficient orders for patients [5]. Based on our data, it can be argued that this was indeed accomplished by using the decision tree for SARS-CoV-2 test ordering to help clinicians navigate the complex test eligibility criteria. However, the implementation of CPOE and CDS systems has been found to provoke strong emotions in care providers, with negative emotions being the most prevalent. In addition to contributing to the stressors that care providers already face, poorly implemented CDSs can fail if they are too cumbersome to be used as intended [18]. A successful CDS system needs to (1) provide clinicians with the best available knowledge when needed, (2) be highly adopted, (3) be effectively used, and (4) result in continuous improvements in knowledge [19].

Evaluating the effective adoption of CDS can be difficult, as care providers always have the option of selecting criteria randomly in order to complete the ordering process. When evaluating the positivity rates for the patient groups that were defined by the decision tree algorithm, we found statistically significant differences (as expected) in rates of SARS-CoV-2 test positivity between asymptomatic and symptomatic patients and between asymptomatic patients without a history of exposure to SARS-CoV-2 and asymptomatic patients with a history of such exposure. These findings suggest that clinicians appropriately used the questions in the CDS algorithm to help triage patients.

Limitations

Our study has several limitations. First, this was an observational study and not a randomized controlled trial. Therefore, other interventions and institutional changes could have explained the decrease in order error rates. Second, the period prior to the implementation of CDS was relatively brief; during this period, a comparatively lower volume of testing was performed. Third, the decision tree was continually modified over time; new indications, such as patients awaiting behavioral health placement, were added relatively late into the pandemic. Some of the positivity rates that were observed in particular patient cohorts could have been influenced by fluctuations in the infection rate within the community.

Conclusions

The leveraging of the EHR and implementation of the decision support algorithm resulted in the decreased incidence of order entry errors, including decreases in the percentage of cancelled and reordered SARS-CoV-2 tests and the rate of duplicate testing, and the appropriate flagging of PUIs. Collectively, these interventions optimized reagent and PPE usage and protected health care workers. The data gathered through the decision tree could be used to predict differences in the likelihood of positive test results for distinct categories of patients, suggesting that clinicians appropriately used the questions in the decision tree algorithm.

Conflicts of Interest

LMF is an unpaid advisory board member for Avsana Labs and has received grant funding for an investigator-initiated study from Biofire Diagnostics.

References

1. Pneumonia of unknown cause – China. World Health Organization. 2020 Jan 05. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229 [accessed 2021-08-16]
2. Zitek T. The appropriate use of testing for COVID-19. West J Emerg Med 2020 Apr 13;21(3):470-472 [FREE Full text] [doi: 10.5811/westjem.2020.4.47370] [Medline: 32302278]
3. Beeching NJ, Fletcher TE, Beadsworth MBJ. Covid-19: testing times. BMJ 2020 Apr 08;369:m1403. [doi: 10.1136/bmj.m1403] [Medline: 32269032]
4. Babiker A, Myers CW, Hill CE, Guarner J. SARS-CoV-2 testing. Am J Clin Pathol 2020 May 05;153(6):706-708 [FREE Full text] [doi: 10.1093/ajcp/aqaa052] [Medline: 32227199]

5. Payne TH, Hoey PJ, Nichol P, Lovis C. Preparation and use of preconstructed orders, order sets, and order menus in a computerized provider order entry system. J Am Med Inform Assoc 2003;10(4):322-329 [FREE Full text] [doi: 10.1197/jamia.M1090] [Medline: 12668686]
6. Horng S, Joseph JW, Calder S, Stevens JP, O'Donoghue AL, Safran C, et al. Assessment of unintentional duplicate orders by emergency department clinicians before and after implementation of a visual aid in the electronic health record ordering system. JAMA Netw Open 2019 Dec 02;2(12):e1916499 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.16499] [Medline: 31790566]
7. Westbrook JI, Li L, Raban MZ, Baysari MT, Mumford V, Prgomet M, et al. Stepped-wedge cluster randomised controlled trial to assess the effectiveness of an electronic medication management system to reduce medication errors, adverse drug events and average length of stay at two paediatric hospitals: a study protocol. BMJ Open 2016 Oct 21;6(10):e011811 [FREE Full text] [doi: 10.1136/bmjopen-2016-011811] [Medline: 27797997]
8. Magid S, Forrer C, Shaha S. Duplicate orders: an unintended consequence of computerized provider/physician order entry (CPOE) implementation: analysis and mitigation strategies. Appl Clin Inform 2012 Oct 17;3(4):377-391 [FREE Full text] [doi: 10.4338/ACI-2012-01-RA-0002] [Medline: 23646085]
9. Wetterneck TB, Walker JM, Blosky MA, Cartmill RS, Hoonakker P, Johnson MA, et al. Factors contributing to an increase in duplicate medication order errors after CPOE implementation. J Am Med Inform Assoc 2011;18(6):774-782 [FREE Full text] [doi: 10.1136/amiajnl-2011-000255] [Medline: 21803925]
10. Tsou AY, Lehmann CU, Michel J, Solomon R, Possanza L, Gandhi T. Safe practices for copy and paste in the EHR. Systematic review, recommendations, and novel model for health IT collaboration. Appl Clin Inform 2017 Jan 11;8(1):12-34 [FREE Full text] [doi: 10.4338/ACI-2016-09-R-0150] [Medline: 28074211]
11. Feketea GM, Vlacha V. A decision-making algorithm for children with suspected coronavirus disease 2019. JAMA Pediatr 2020 Dec 01;174(12):1220-1222. [doi: 10.1001/jamapediatrics.2020.2999] [Medline: 32955559]
12. 2020 key metrics. Children's Health. URL: https://www.childrens.com/footer/about/our-system/key-metrics [accessed 2021-08-16]
13. Schuchat A, CDC COVID-19 Response Team. Public health response to the initiation and spread of pandemic COVID-19 in the United States, February 24-April 21, 2020. MMWR Morb Mortal Wkly Rep 2020 May 08;69(18):551-556 [FREE Full text] [doi: 10.15585/mmwr.mm6918e2] [Medline: 32379733]
14. Oster AM, Kang GJ, Cha AE, Beresovsky V, Rose CE, Rainisch G, et al. Trends in number and distribution of COVID-19 hotspot counties - United States, March 8-July 15, 2020. MMWR Morb Mortal Wkly Rep 2020 Aug 21;69(33):1127-1132 [FREE Full text] [doi: 10.15585/mmwr.mm6933e2] [Medline: 32817606]
15. Myers JF, Snyder RE, Porse CC, Tecle S, Lowenthal P, Danforth ME, Traveler Monitoring Team. Identification and monitoring of international travelers during the initial phase of an outbreak of COVID-19 - California, February 3-March 17, 2020. MMWR Morb Mortal Wkly Rep 2020 May 15;69(19):599-602 [FREE Full text] [doi: 10.15585/mmwr.mm6919e4] [Medline: 32407299]
16. McGovern OL, Stenger M, Oliver SE, Anderson TC, Isenhour C, Mauldin MR, et al. Demographic, clinical, and epidemiologic characteristics of persons under investigation for coronavirus disease 2019-United States, January 17-February 29, 2020. PLoS One 2021 Apr 15;16(4):e0249901. [doi: 10.1371/journal.pone.0249901] [Medline: 33857209]
17. Updated guidance on evaluating and testing persons for coronavirus disease 2019 (COVID-19). Centers for Disease Control and Prevention. URL: https://emergency.cdc.gov/han/2020/han00429.asp [accessed 2021-08-16]
18. Sittig DF, Krall M, Kaalaas-Sittig J, Ash JS. Emotional aspects of computer-based provider order entry: a qualitative study. J Am Med Inform Assoc 2005;12(5):561-567 [FREE Full text] [doi: 10.1197/jamia.M1711] [Medline: 15905478]
19. 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(Suppl 1):S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]

Abbreviations

AGP: aerosol-generating procedure
CDS: clinical decision support
CPOE: computerized provider order entry
EHR: electronic health record
PPE: personal protective equipment
PUI: person under investigation

Edited by G Eysenbach; submitted 22.07.21; peer-reviewed by J Walsh; comments to author 13.08.21; revised version received 18.08.21; accepted 19.09.21; published 18.10.21.

Please cite as:
Luu HS, Filkins LM, Park JY, Rakheja D, Tweed J, Menzies C, Wang VJ, Mittal V, Lehmann CU, Sebert ME
Harnessing the Electronic Health Record and Computerized Provider Order Entry Data for Resource Management During the COVID-19 Pandemic: Development of a Decision Tree
JMIR Med Inform 2021;9(10):e32303
URL: https://medinform.jmir.org/2021/10/e32303
doi:10.2196/32303
PMID:34546942

Original Paper

Verifying the Feasibility of Implementing Semantic Interoperability in Different Countries Based on the OpenEHR Approach: Comparative Study of Acute Coronary Syndrome Registries

Lingtong Min[1], PhD; Koray Atalag[2], MD, PhD; Qi Tian[3], PhD; Yani Chen[3], PhD; Xudong Lu[3], PhD

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China

[2]Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand

[3]College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, China

Corresponding Author:
Xudong Lu, PhD
College of Biomedical Engineering & Instrument Science
Zhejiang University
Room 512, Zhouyiqing Building
38 Zheda Road
Hangzhou
China
Phone: 86 13957118891
Email: lvxd@zju.edu.cn

Abstract

Background: The semantic interoperability of health care information has been a critical challenge in medical informatics and has influenced the integration, sharing, analysis, and use of medical big data. International standard organizations have developed standards, approaches, and models to improve and implement semantic interoperability. The openEHR approach—one of the standout semantic interoperability approaches—has been implemented worldwide to improve semantic interoperability based on reused archetypes.

Objective: This study aimed to verify the feasibility of implementing semantic interoperability in different countries by comparing the openEHR-based information models of 2 acute coronary syndrome (ACS) registries from China and New Zealand.

Methods: A semantic archetype comparison method was proposed to determine the semantics reuse degree of reused archetypes in 2 ACS-related clinical registries from 2 countries. This method involved (1) determining the scope of reused archetypes; (2) identifying corresponding data items within corresponding archetypes; (3) comparing the semantics of corresponding data items; and (4) calculating the number of mappings in corresponding data items and analyzing results.

Results: Among the related archetypes in the two ACS-related, openEHR-based clinical registries from China and New Zealand, there were 8 pairs of reusable archetypes, which included 89 pairs of corresponding data items and 120 noncorresponding data items. Of the 89 corresponding data item pairs, 87 pairs (98%) were mappable and therefore supported semantic interoperability, and 71 pairs (80%) were labeled as “direct mapping” data items. Of the 120 noncorresponding data items, 114 (95%) data items were generated via archetype evolution, and 6 (5%) data items were generated via archetype localization.

Conclusions: The results of the semantic comparison between the two ACS-related clinical registries prove the feasibility of establishing the semantic interoperability of health care data from different countries based on the openEHR approach. Archetype reuse provides data on the degree to which semantic interoperability exists when using the openEHR approach. Although the openEHR community has effectively promoted archetype reuse and semantic interoperability by providing archetype modeling methods, tools, model repositories, and archetype design patterns, the uncontrolled evolution of archetypes and inconsistent localization have resulted in major challenges for achieving higher levels of semantic interoperability.

(JMIR Med Inform 2021;9(10):e31288) doi:10.2196/31288

KEYWORDS

semantic interoperability; openEHR; archetype; registry; acute coronary syndrome

Introduction

Due to the rapid development of information and communication technologies and their continuous application in the medical domain, massive electronic medical data and information have been generated by medical information systems, services, and devices. These vast amounts of medical data and information have the potential to improve the safety and quality of medical services, reduce the cost of such services, and promote medical research [1]. The realization of this potential needs to be supported by the effective application of advanced information and communication technologies that manage medical big data, such as big data analysis technologies and artificial intelligence technologies. However, the effective application of these technologies is premised on implementing semantic interoperability, which is an indispensable basis for the construction, sharing, analysis, and use of medical big data. The semantic interoperability of clinical data and information has been a critical challenge and hot topic in medical informatics [2].

To implement and improve semantic interoperability, international standard organizations have developed approaches for constructing standards based on multilevel medical information models. Such organizations include Health Level Seven (HL7) International [3], the International Organization for Standardization (ISO) [4], and the openEHR Foundation [5].

To promote the exchange and sharing of medical information, HL7 International has developed a series of standards based on the HL7 development framework and medical information models, including the HL7 v2 and v3 messaging standards and HL7 Clinical Document Architecture standards. These standards and models have made important contributions to the improvement of semantic interoperability, especially semantic interoperability in the electronic exchange of messages and medical documents [6,7]. The HL7 Clinical Document Architecture standards have become core standards for the interoperability of medical documents. Although HL7 v3 can improve semantic interoperability, the inconsistency and complexity of its reference information model and modeling approach have limited its implementation [8].

To solve the complexity problem, HL7 International launched the Fast Healthcare Interoperability Resources (FHIR) standard and a limited set of information models to easily implement semantic interoperability [9]. FHIR provide a set of resources for expressing the structure and semantics of the most commonly exchanged information and provide a flexible extension mechanism for defining the semantics of information that is not included in FHIR. The feasibility of using FHIR to improve the semantic interoperability of medical information systems and applications has been verified [10-13].

ISO13606 is a series of health informatics standards that facilitate electronic health record (EHR) data exchange and communication between EHR systems and data repositories [14]. ISO13606 is based on the openEHR 2-level modeling method, which involves a simplified reference model, and is regarded as a subset of the openEHR specifications [15].

The openEHR Foundation provides a comprehensive information architecture for representing the entirety of EHR content in a structured and interoperable form. This architecture includes a modeling framework that uses domain-specific languages and intuitive modeling tools. The openEHR Foundation tries to achieve sustainable semantic interoperability by using a consistent multilevel modeling method to define reusable domain concepts and their formal semantics [16].

The models within the openEHR approach consist of a reference model, archetypes, and templates. The reference model is a stable information model that defines a logical medical information architecture and includes a demographic information model, an EHR information model, the EHR Extract Information model, data types, and data structures. An archetype is a reusable information model that comprises a maximal set of content element definitions and general constraints. Templates are context-specific data set definitions that are created by combining and constraining relevant archetypes for generating forms, documents, data persistence, and messages. Archetypes and templates are defined by domain experts using a formal editorial and publication process to foster semantic interoperability.

OpenEHR-related research is gradually becoming one of the most discussed semantic interoperability–related research topics. Such research involves archetype modeling [17-23], data persistence [24-26], language design [27], model mapping [28], model retrieval [29,30], and reuse [19].

The openEHR approach has been adopted in many countries to improve the semantic interoperability of medical information systems [31], such as those in Norway [30], China [17], Portugal [32], Brazil [15,33], and Germany [34]. Moreover, the feasibility of improving semantic interoperability based on the openEHR approach has been verified in many domains, including genomics [21], clinical decision support systems [34], clinical registries [35], clinical data sets [36], and EHRs [37].

To the best of our knowledge, there has been no research on verifying the feasibility of implementing semantic interoperability in different countries based on the openEHR approach.

Methods

Study Design

In order to verify the feasibility and degree of implementing semantic interoperability in 2 countries based on the openEHR approach, we conducted a semantic comparison of the archetypes in 2 acute coronary syndrome (ACS)–related clinical data registries from China and New Zealand.

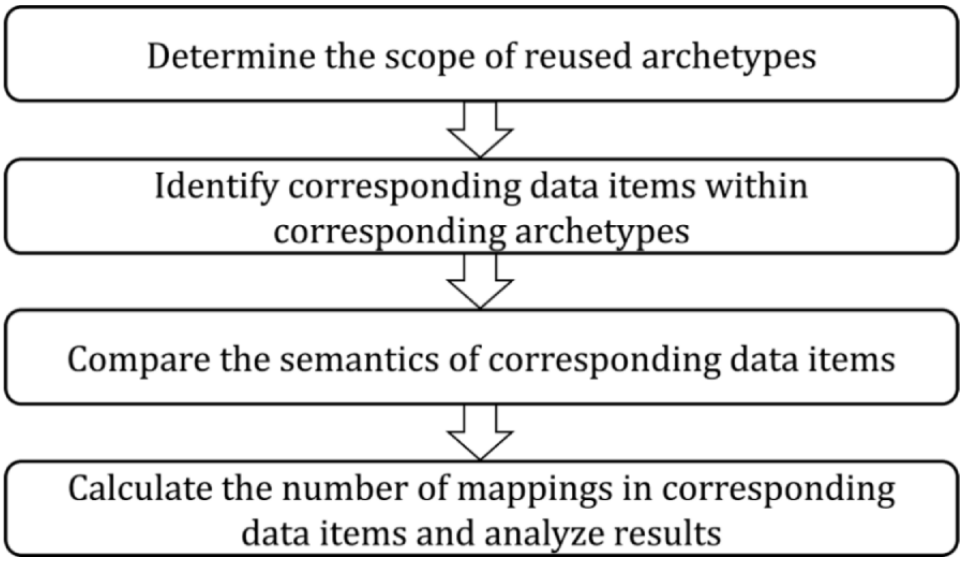
In this paper, the analyzed ACS-related data registry from China was the Coronary Computed Tomography Angiography (CCTA) registry [35], which is used to improve the early detection of coronary atherosclerosis as part of a national key research and development project. The CCTA registry was designed to collect clinical data from patients who have undergone CCTA examination. The analyzed ACS-related registry from New Zealand was the All New Zealand ACS Quality Improvement

(ANZACS-QI) program. The ANZACS-QI registry is a cardiac registry for ACS events and cardiac procedures. The data management of these two registries was based on the openEHR approach [36].

Archetype reuse is key for establishing semantic interoperability based on the openEHR approach. Therefore, the semantic expression comparison of reused archetypes can be conducted to determine the degree to which different openEHR-based clinical systems and applications achieve semantic interoperability.

This paper proposes a semantic archetype comparison method for determining the degree of semantic interoperability via archetype reuse in 2 ACS-related clinical registries from 2 countries. This semantic archetype comparison method consisted of the following four steps: (1) determining the scope of reused archetypes; (2) identifying corresponding data items within corresponding archetypes; (3) comparing the semantics of corresponding data items; and (4) calculating the number of mappings in corresponding data items and analyzing results, as shown in Figure 1.

Figure 1. The procedure for the semantic archetype comparison method.



Determining the Scope of Reused Archetypes

The scope of reused archetypes was determined based on the comparison of archetype names, types, descriptions, and metadata. An archetype name is an archetype's unique identifier, and it reflects an archetype's semantics, including the archetype type, corresponding domain concept, and version information. The types of archetypes include the observation, instruction, action, evaluation, and administrative archetypes.

The continuous improvement of archetypes over time results in changes in the archetype version based on certain rules. Localizing existing archetypes also leads to changes in archetype names, which usually manifest as the addition of suffixes to archetype names. Therefore, while determining the scope of reusable archetypes, if the names of 2 archetypes were the same or if there was a localization relationship between 2 archetypes, they were matched for reuse.

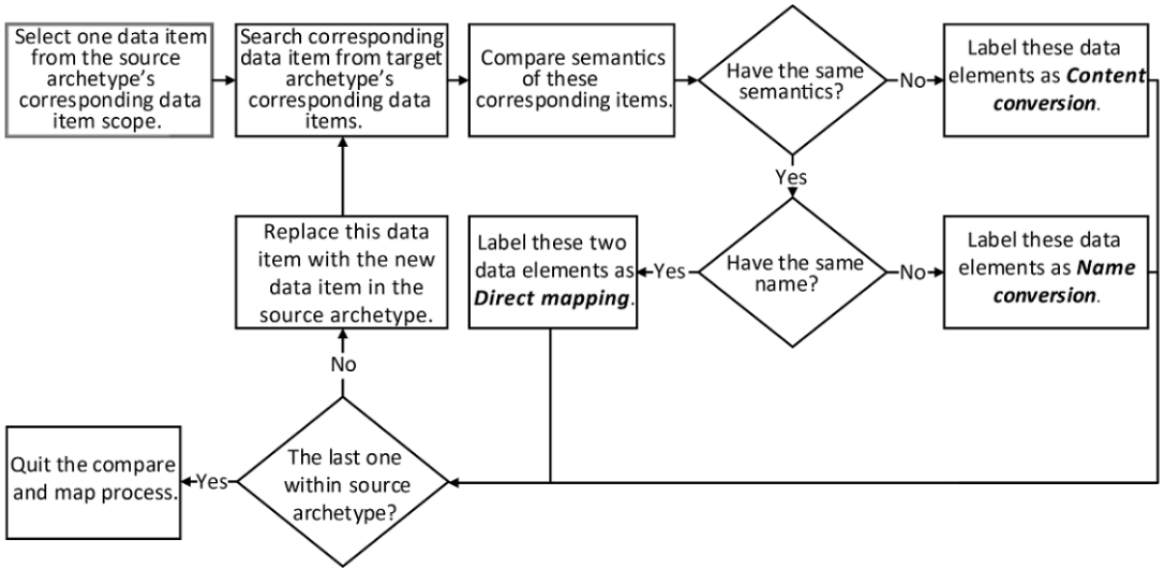
Identifying Corresponding Data Items Within Corresponding Archetypes

It was necessary to identify the corresponding data items in each pair of corresponding reused archetypes. This was done by comparing the data items' semantic descriptions. If the semantic descriptions of 2 data items were the same or similar, they were marked as corresponding data items; otherwise, they were marked as noncorresponding data items. Noncorresponding data items indicated semantic gaps or differences among the data items from reused archetypes that may hinder semantic interoperability.

Comparing the Semantics of Corresponding Data Items

The semantic comparison of the corresponding data items within reused archetypes mainly involved data element names and semantic descriptions. In accordance with this procedure, the corresponding data items within reused archetypes were labeled as "direct mapping," "name conversion," and "content conversion" data items based on the mapping results, as shown in Figure 2. These semantic mapping–based data item types were used to determine the degree to which an archetype's data items supported semantic interoperability. The "direct mapping" corresponding data items could support complete semantic interoperability without human intervention. The "name conversion" corresponding data items could support semantic interoperability via manual or automated name mapping. The "content conversion" corresponding data items referred to items that made it challenging to achieve semantic interoperability.

Figure 2. The process of the semantic archetype comparison.



Calculating the Number of Mappings in Corresponding Data Items and Analyzing Results

First, the number of mappings for each data item type among the corresponding data items in each reused archetype and in all reused archetypes was calculated. Second, the direct mapping ratio and the mappable ratio of the corresponding data items in each reused archetype and in all reused archetypes were calculated. Finally, the data obtained from the abovementioned calculations were analyzed to illustrate the reusability of related archetypes from the two registries.

Results

To illustrate the feasibility of implementing semantic interoperability between the two coronary clinical data registries based on the openEHR approach, we compared the corresponding data elements within reused archetypes by using statistical analyses.

The Scope of Reused Archetypes and Corresponding Data Items

By comparing the ACS-related archetypes in the two registries, 8 pairs of reused archetypes were identified. These are shown in Textbox 1.

Textbox 1. Corresponding acute coronary syndrome–related reused archetypes from China and New Zealand.

Reused archetypes from China

1. openEHR-EHR-EVALUATION.problem_diagnosis.v1
2. openEHR-EHR-ACTION.medication.v1
3. openEHR-EHR-OBSERVATION.imaging_exam_result.v0
4. openEHR-EHR-OBSERVATION.blood_pressure.v2
5. openEHR-EHR-EVALUATION.tobacco_smoking_summary.v1
6. openEHR-EHR-EVALUATION.family_history.v2
7. openEHR-EHR-OBSERVATION.body_weight.v2
8. openEHR-EHR-OBSERVATION.height.v2

Reused archetypes from New Zealand that corresponded with those from China

1. openEHR-EHR-EVALUATION.problem_diagnosis_nehta.v1
2. openEHR-EHR-ACTION.medication.v1
3. openEHR-EHR-OBSERVATION.imaging_exam.v1
4. openEHR-EHR-OBSERVATION.blood_pressure.v1
5. openEHR-EHR-EVALUATION.tobacco_use_summary.v1
6. openEHR-EHR-EVALUATION.family_history.v1
7. openEHR-EHR-OBSERVATION.body_weight.v1
8. openEHR-EHR-OBSERVATION.height.v1

The semantic comparison of data items in the reused archetypes revealed 89 pairs of corresponding data items and 120 noncorresponding data items from the two countries. Of the 120 noncorresponding data items, 86 were from the reused archetypes in China, and 34 were from the reused archetypes in New Zealand. Further, 50.9% (89/175) of the data items from China were corresponding data items, and 49.1% (86/175) of the data items were noncorresponding data items. Additionally, 72.4% (89/123) of the data items from New Zealand were corresponding data items, and 27.6% (34/123) of the data items were noncorresponding data items.

The reasons for and the proportion distribution of the noncorresponding data items in the reused archetypes were compared and analyzed. Of the 86 noncorresponding items from China, 83 were generated via archetype evolution and 3 were generated via archetype localization. Further, 31 of the 34 noncorresponding data items from New Zealand were generated via archetype version evolution and 3 were generated via archetype localization.

Results of the Semantic Comparison of Corresponding Data Items

The results of the semantic comparison of corresponding data items in reused archetypes are shown in the Table 1.

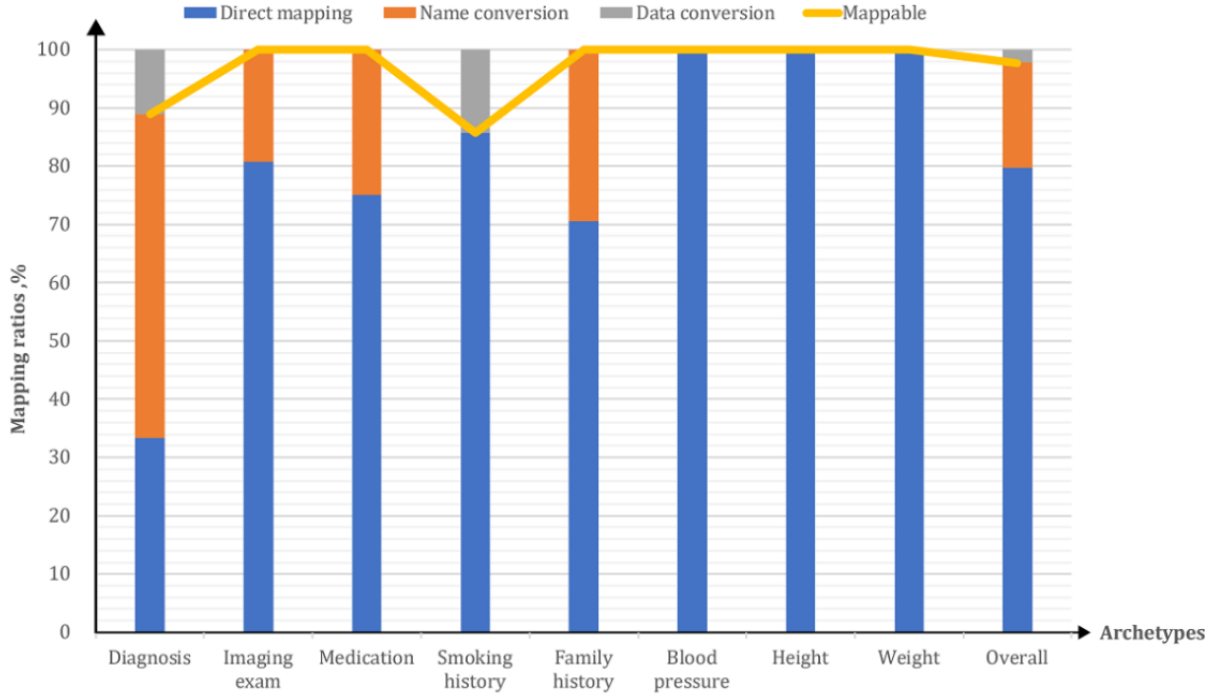
Table 1. The results of the semantic comparison of corresponding data items in reused archetypes.

Domain concept	Total number of data items	Direct mapping data items, n	Name conversion data items, n	Content conversion data items, n	Direct mapping ratio, %	Mappable ratio, %
Diagnosis	9	3	5	1	33	89
Imaging exam	26	21	5	0	81	100
Medication	4	3	1	0	75	100
Blood pressure	16	16	0	0	100	100
Smoking history	7	6	0	1	14	86
Family history	17	12	5	0	71	100
Weight	5	5	0	0	100	100
Height	5	5	0	0	100	100
All concepts	89	71	16	2	80	98

“Direct mapping” data items referred to data items in reused archetypes that could support semantic interoperability without any modification. The semantics of the “name conversion” data items were the same, but the names of the data items were different (eg, “Problem/Diagnosis” vs “Problem/Diagnosis name” for the corresponding reused “diagnosis” archetypes). “Name conversion” data items referred to corresponding data items that could also support semantic interoperability, but some additional conversion mapping would be required. However, “content conversion” data items had the problem of incomplete semantic matching between data items and therefore limited the realization of semantic interoperability. As such, the “mappable ratio” of data items that supported semantic interoperability included the “direct mapping” and “name conversion” data items.

The mapping ratios of the corresponding data items of the reused archetypes are shown in Figure 3. The mappable proportion of all corresponding data items within reused archetypes was as high as 98% (87/89). Except for the mapping ratios of the “diagnosis” (8/9, 89%) and “smoking history” (6/7, 86%) archetypes, the mapping ratios of the corresponding items of all archetypes were 100%.

Figure 3. The various mapping ratios of the corresponding data items of reused archetypes.



Discussion

Principal Findings

Archetype reuse is the main determinant of the semantic interoperability of openEHR-based medical information. This study attempted to verify the feasibility of achieving semantic interoperability and the extent to which the semantic interoperability of clinical information can be achieved by assessing archetype reuse in 2 ACS registries from China and New Zealand. The results of our novel semantic comparison method indicated that the ratio of direct mappings between the corresponding data items of reused archetypes reached 98% (87/89), and of the 8 reused archetypes, 6 (75%) had data items that achieved a direct mapping ratio of 100%. This very high degree of reused archetype direct mapping proves the feasibility of achieving semantic interoperability by using the openEHR approach.

Although the two clinical data registries for China and New Zealand both contained data about ACS, they have different scopes due to their different purposes. The ANZACS-QI registry is used to support clinical quality improvement and research, whereas the CCTA registry is used to improve the early detection of coronary atherosclerosis. Therefore, the aspects that are important for New Zealand may not be as relevant to China and vice versa. For example, invasive management and emergency management are two essential parts of ACS management in New Zealand; however, the CCTA registry barely covers these aspects. Conversely, CCTA examination is the cornerstone aspect of the CCTA registry; however, this is not relevant to the ANZACS-QI registry.

For the past 2 decades, extensive research on achieving semantic interoperability through archetype reuse has been undertaken by the openEHR community. Further, valuable insights have been gained through the evaluation of real-world deployments. Although the flexibility and extensibility of the openEHR approach allow for a high degree of freedom when creating and modifying archetypes for specific purposes, it is critically important to ensure that these archetypes can be reused. To this end, extensive efforts and resources have been devoted to the development and sharing of archetypes at an international level. The development of volunteer-created editorial processes and the Clinical Knowledge Manager (CKM) model repository [38], which is freely available to the global openEHR community, were significant steps toward creating more than 500 archetypes that represent a substantial portion of EHRs. These efforts have provided an essential foundation for archetype reuse among various applications worldwide.

Although these efforts have drawn remarkable attention and have resulted in significant breakthroughs, there are still important challenges to archetype reuse among various systems and applications [5,17,19,22,23]. These challenges include (1) the uncontrolled evolution of existing archetypes across various versions of the CKM; (2) inconsistencies in new archetype development; (3) unanticipated semantic changes generated in successive versions of an archetype; and (4) inconsistencies in localization and terminology bindings.

We observed several of these issues in our study. For example, the semantic mismatch between the "openEHR-EHR-EVALUATION.problem_diagnosis.v1" archetype in the international CKM and the "openEHR-EHR-EVALUATION.problem_diagnosis_nehta.v1" archetype in the Australian National E-Health Transition Authority version of the CKM resulted in the inability to directly map all data elements. Another example is the use of two different versions of archetypes—the "openEHR-EHR-EVALUATION.family_history.v2" archetype and the "openEHR-EHR-EVALUATION.family_history.v1" archetype in the two modeling registries.

From a methodological point of view, a graphical archetype discovery method [29] and formal archetype modeling methods [17,22,23,35,39] for facilitating robust archetype development have previously been proposed and validated. Although these efforts could improve archetype reuse, achieving real-world semantic interoperability, as illustrated in this study, between 2 ACS registries largely depends on human factors. Therefore, a prescriptive modeling framework is needed. Integrating semantic-based model discovery and automated modeling assistance services into archetype development tools is a potential solution for meeting this need. However, to the best of our knowledge, such methodologies and tools do not exist and warrant future research. In addition, the establishment of archetype update notification services, which would notify developers of information systems (ie, those that use existing archetypes) about changes, can be very helpful for ensuring the continuity of semantic interoperability over time as new changes are incorporated.

This *Discussion* section would be incomplete if we did not highlight the parallels between openEHR archetypes and HL7 FHIR, as both are formal models of health care information. The comparison of the FHIR and openEHR approaches is shown in the Table 2. On one hand, while anyone can create new archetypes that meet specific requirements, FHIR do not allow for the creation of new resources but provide extensions and profiling mechanisms for customizing existing resources. On the other hand, the openEHR multilevel modeling method, which is based on a stable reference model, provides a well-defined framework for expressing complex concepts and adapts to the rapid evolution of domain concepts better than the FHIR approach. Although the centrally developed and published FHIR are concrete and immutable (ie, factors that foster model reuse), the unbounded extension mechanism still relies on human factors for aligning resources with existing extensions. Ultimately, the compromise between the freedom of creating new information models and the need to ensure semantic alignment at a global scale applies equally to both the openEHR and FHIR formalisms. Therefore, our methodology and results should also be applicable to FHIR.

Table 2. The comparison between the Fast Healthcare Interoperability Resources (FHIR) and openEHR approaches.

Comparison aspects	FHIR approach	OpenEHR approach
Scope and purpose	FHIR is a new Health Level Seven specification for defining the structure and semantics of health care information that is involved in health information exchange. It is not engineered for persistence and the modeling of all electronic health record data.	OpenEHR defines the structure and semantics of all health information in electronic health records and is engineered for persistence and health information exchange.
Reference model	The model does not have a separate layer and uses a mix of data types and structural resources.	The model has a discrete and stable layer that comprises building blocks, upon which archetypes are built.
Information model	FHIR	OpenEHR archetypes
Composition and constraints	FHIR profile	OpenEHR template
Extensibility	FHIR extensions that are discoverable via Uniform Resource Identifiers	New archetypes or archetype specializations for adding new data items, value sets, and tighter constraints
Localization	Localization is not well defined. Localization is possible with extensions, but this is under review.	Localization is a well-defined section in archetypes for data elements and value sets.
Terminology support	Supports terminology bindings	Supports terminology bindings
Reference relationship	The resources may refer to those outside of FHIR.	The archetypes allow for the linking of other archetypes via an archetype slotting mechanism.

Due to the ever-increasing amounts of clinical domain knowledge and various local implementation needs, the consistency and reusability of new or modified archetypes are grand challenges to implementing semantic interoperability. In our study, we were fortunate, as the number of these inconsistencies were minimal. As such, we were able to demonstrate the feasibility of achieving a very high level of semantic interoperability between 2 clinical registries from 2 very distinct countries.

Conclusions

The feasibility of implementing semantic interoperability in 2 ACS registries that are based on the openEHR approach was verified by the results of our semantic comparison of reused archetypes. Continuous improvement and localized archetype modification may reduce the proportion of direct mappings among data items in reused archetypes and result in a gap between actual semantic interoperability and theoretical semantic interoperability. Although the openEHR community has provided an essential foundation for archetype reuse through robust editorial processes and a freely available CKM, there are still important challenges to archetype reuse.

Acknowledgments

This study was funded by the Chinese National Science and Technology Major Project (grant 2016YFC1300303).

Authors' Contributions

LM proposed the study idea, drafted the manuscript, and designed the archetype mapping method. LM, QT, and YC performed the archetype mapping procedure for the two different acute coronary syndrome clinical data registries of China and New Zealand. KA led the openEHR modeling of the All New Zealand Acute Coronary Syndrome Quality Improvement registry and provided the archetypes and templates for the comparative analysis. KA and XL made critical revisions and improvements to the manuscript.

Conflicts of Interest

None declared.

References

1. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract 2017 Mar;36(1):3-11 [FREE Full text] [doi: 10.23876/j.krcp.2017.36.1.3] [Medline: 28392994]
2. Adel E, El-Sappagh S, Barakat S, Elmogy M. Chapter 13 - Ontology-based electronic health record semantic interoperability: A survey. In: Dey N, Ashour AS, Fong SJ, Borra S, editors. U-Healthcare Monitoring Systems Volume 1: Design and Applications. Amsterdam, The Netherlands: Elsevier; 2019:315-352.
3. Introduction to HL7 standards. Health Level Seven International. URL: https://www.hl7.org/implement/standards/index.cfm?ref=nav [accessed 2021-01-15]
4. Standards by ISO/TC 215: Health informatics. International Organization for Standardization. URL: https://www.iso.org/committee/54960/x/catalogue/ [accessed 2021-01-15]
5. Clinical models program. openEHR Foundation. URL: https://www.openehr.org/programs/clinicalmodels/ [accessed 2021-01-15]
6. Vida M, Lupse O, Stoicu-Tivadar L. Improving the interoperability of healthcare information systems through HL7 CDA and CCD standards. 2012 Presented at: 2012 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)Informatics (SACI); May ; Timisoara, Romania. IEEE; May 24-26, 2012; Timisoara, Romania p. 157-161. [doi: 10.1109/saci.2012.6249994]
7. Porrasmaa J, Mykkänen J, Tarhonen T, Jalonen M, Kemppainen P, Ensio A, et al. Application of HL7 CDA R2 and V3 messaging for national ePrescription in Finland. 2008 Presented at: 9th International HL7 Interoperability Conference (IHIC); October 8-11, 2008; Hersonissos, Crete, Greece.
8. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR, Fourth Edition. Switzerland: Springer International Publishing; 2016.
9. FHIR overview. HL7 FHIR. URL: https://www.hl7.org/fhir/overview.html [accessed 2021-01-15]
10. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019 Jun;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
11. Leroux H, Metke-Jimenez A, Lawley MJ. Towards achieving semantic interoperability of clinical study data with FHIR. J Biomed Semantics 2017 Sep 19;8(1):41 [FREE Full text] [doi: 10.1186/s13326-017-0148-7] [Medline: 28927443]
12. Gruendner J, Wolf N, Tögel L, Haller F, Prokosch H, Christoph J. Integrating genomics and clinical data for statistical analysis by using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): System design and implementation. J Med Internet Res 2020 Oct 07;22(10):e19879 [FREE Full text] [doi: 10.2196/19879] [Medline: 33026356]
13. 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]
14. ISO 13606-1:2019 - Health informatics — Electronic health record communication — Part 1: Reference model. International Organization for Standardization. 2019. URL: https://www.iso.org/standard/67868.html [accessed 2021-01-15]
15. Pahl C, Zare M, Nilashi M, de Faria Borges MA, Weingaertner D, Detschew V, et al. Role of OpenEHR as an open source solution for the regional modelling of patient data in obstetrics. J Biomed Inform 2015 Jun;55:174-187 [FREE Full text] [doi: 10.1016/j.jbi.2015.04.004] [Medline: 25900270]
16. What is openEHR? openEHR Foundation. URL: https://www.openehr.org/about/what_is_openehr [accessed 2021-01-16]
17. Min L, Tian Q, Lu X, Duan H. Modeling EHR with the openEHR approach: an exploratory study in China. BMC Med Inform Decis Mak 2018 Aug 29;18(1):75 [FREE Full text] [doi: 10.1186/s12911-018-0650-6] [Medline: 30157838]
18. Wei P, Atalag K, Day K. An openEHR approach to detailed clinical model development: Tobacco smoking summary archetype as a case study. Appl Clin Inform 2019 Mar;10(2):219-228 [FREE Full text] [doi: 10.1055/s-0039-1681074] [Medline: 30919398]
19. Leslie H. OpenEHR archetype use and reuse within multilingual clinical data sets: Case study. J Med Internet Res 2020 Nov 02;22(11):e23361 [FREE Full text] [doi: 10.2196/23361] [Medline: 33035176]
20. Li M, Leslie H, Qi B, Nan S, Feng H, Cai H, et al. Development of an openEHR template for COVID-19 based on clinical guidelines. J Med Internet Res 2020 Jun 10;22(6):e20239 [FREE Full text] [doi: 10.2196/20239] [Medline: 32496207]
21. Mascia C, Uva P, Leo S, Zanetti G. OpenEHR modeling for genomics in clinical practice. Int J Med Inform 2018 Dec;120:147-156 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.10.007] [Medline: 30409340]

22. Moreno-Conde A, Moner D, da Cruz WD, Santos MR, Maldonado JA, Robles M, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc 2015 Jul;22(4):925-934 [FREE Full text] [doi: 10.1093/jamia/ocv008] [Medline: 25796595]
23. Moner D, Maldonado JA, Robles M. Archetype modeling methodology. J Biomed Inform 2018 Mar;79:71-81 [FREE Full text] [doi: 10.1016/j.jbi.2018.02.003] [Medline: 29454107]
24. Helou SE, Kobayashi S, Yamamoto G, Kume N, Kondoh E, Hiragi S, et al. Graph databases for openEHR clinical repositories. International Journal of Computational Science and Engineering 2019;20(3):281. [doi: 10.1504/ijcse.2019.103955]
25. Kalogiannis S, Deltouzos K, Zacharaki EI, Vasilakis A, Moustakas K, Ellul J, et al. Integrating an openEHR-based personalized virtual model for the ageing population within HBase. BMC Med Inform Decis Mak 2019 Jan 28;19(1):25 [FREE Full text] [doi: 10.1186/s12911-019-0745-8] [Medline: 30691467]
26. Wang L, Min L, Wang R, Lu X, Duan H. Archetype relational mapping - a practical openEHR persistence solution. BMC Med Inform Decis Mak 2015 Nov 05;15(1):88 [FREE Full text] [doi: 10.1186/s12911-015-0212-0] [Medline: 26541142]
27. Tian Q, Han Z, An J, Lu X, Duan H. Representing rules for clinical data quality assessment based on openEHR guideline definition language. Stud Health Technol Inform 2019 Aug 21;264:1606-1607. [doi: 10.3233/SHTI190557] [Medline: 31438254]
28. Costa CM, Menárguez-Tortosa M, Fernández-Breis JT. Clinical data interoperability based on archetype transformation. J Biomed Inform 2011 Oct;44(5):869-880 [FREE Full text] [doi: 10.1016/j.jbi.2011.05.006] [Medline: 21645637]
29. Yang L, Huang X, Li J. Discovering clinical information models online to promote interoperability of electronic health records: A feasibility study of openEHR. J Med Internet Res 2019 May 28;21(5):e13504 [FREE Full text] [doi: 10.2196/13504] [Medline: 31140433]
30. Christensen B, Ellingsen G. Evaluating model-driven development for large-scale EHRs through the openEHR approach. Int J Med Inform 2016 May;89:43-54 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.02.004] [Medline: 26980358]
31. OpenEHR deployed solutions. openEHR Foundation. URL: https://www.openehr.org/openehr_in_use/deployed_solutions/ [accessed 2021-01-16]
32. Hak F, Oliveira D, Abreu N, Leuschner P, Abelha A, Santos M. An openEHR adoption in a Portuguese healthcare facility. Procedia Comput Sci 2020;170:1047-1052 [FREE Full text] [doi: 10.1016/j.procs.2020.03.075]
33. Pellison FC, Rijo RPCL, Lima VC, Crepaldi NY, Bernardi FA, Galliez RM, et al. Data integration in the Brazilian public health system for tuberculosis: Use of the semantic web to establish interoperability. JMIR Med Inform 2020 Jul 06;8(7):e17176 [FREE Full text] [doi: 10.2196/17176] [Medline: 32628611]
34. Wulff A, Haarbrandt B, Tute E, Marschollek M, Beerbaum P, Jack T. An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. Artif Intell Med 2018 Jul;89:10-23 [FREE Full text] [doi: 10.1016/j.artmed.2018.04.012] [Medline: 29753616]
35. Min L, Tian Q, Lu X, An J, Duan H. An openEHR based approach to improve the semantic interoperability of clinical data registry. BMC Med Inform Decis Mak 2018 Mar 22;18(Suppl 1):15 [FREE Full text] [doi: 10.1186/s12911-018-0596-8] [Medline: 29589572]
36. Atalag K, Farris A, Kerr A. Information modelling of ANZACS-QI datasets to support data management using openEHR. 2015 Presented at: 14th Annual Health Informatics New Zealand Conference (HINZ 2015); June 22-26, 2015; Christchurch, New Zealand.
37. Oliveira D, Miranda FMM, Coimbra A, Abreu N, Leuschner P, Abelha AC. OpenEHR meets interoperability and knowledge engineering. International Journal of Reliable and Quality E-Healthcare 2020;9(1):1-12. [doi: 10.4018/IJRQEH.2020010101]
38. Clinical knowledge manager. openEHR Foundation. URL: https://ckm.openehr.org/ckm/ [accessed 2021-01-24]
39. Leslie H. Archetype design patterns. openEHR wiki. URL: https://openehr.atlassian.net/wiki/spaces/healthmod/pages/90507705/Archetype+Design+Patterns [accessed 2021-01-22]

Abbreviations

ACS: acute coronary syndrome
ANZACS-QI: All New Zealand Acute Coronary Syndrome Quality Improvement
CCTA: Coronary Computed Tomography Angiography
CKM: Clinical Knowledge Manager
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven
ISO: International Organization for Standardization

Edited by G Eysenbach; submitted 16.06.21; peer-reviewed by S Kobayashi, N Deng; comments to author 08.07.21; revised version received 19.07.21; accepted 01.08.21; published 19.10.21.

Please cite as:
Min L, Atalag K, Tian Q, Chen Y, Lu X
Verifying the Feasibility of Implementing Semantic Interoperability in Different Countries Based on the OpenEHR Approach: Comparative Study of Acute Coronary Syndrome Registries
JMIR Med Inform 2021;9(10):e31288
URL: https://medinform.jmir.org/2021/10/e31288
doi:10.2196/31288
PMID:34665150

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org