Received: 17 July 2020 | Revised: 21 October 2020 | Accepted: 2 January 2021

DOI: 10.1002/wics.1549

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Challenges and opportunities beyond structured data in analysis of electronic health records

Maryam Tayefi[1] | **Phuong Ngo**[1] | **Taridzo Chomutare**[1] | **Hercules Dalianis**[1,2] | **Elisa Salvi**[1] | **Andrius Budrionis**[1] | **Fred Godtliebsen**[1,3]

[1]Norwegian Centre for E-health Research, Tromsø, Norway

[2]Stockholm University, Stockholm, Sweden

[3]UiT The Arctic University of Norway, Tromsø, Norway

Correspondence
Maryam Tayefi, Norwegian Centre for E-health Research, Tromsø, Norway.
Email: maryam.tayefi@ehealthresearch.no

Funding information
Helse Nord RHF, Grant/Award Number: HNF1395-18; Tromsø Forskningsstiftelse (Tromsø Research Foundation), Grant/Award Number: A33027

Abstract

Electronic health records (EHR) contain a lot of valuable information about individual patients and the whole population. Besides structured data, unstructured data in EHRs can provide extra, valuable information but the analytics processes are complex, time-consuming, and often require excessive manual effort. Among unstructured data, clinical text and images are the two most popular and important sources of information. Advanced statistical algorithms in natural language processing, machine learning, deep learning, and radiomics have increasingly been used for analyzing clinical text and images. Although there exist many challenges that have not been fully addressed, which can hinder the use of unstructured data, there are clear opportunities for well-designed diagnosis and decision support tools that efficiently incorporate both structured and unstructured data for extracting useful information and provide better outcomes. However, access to clinical data is still very restricted due to data sensitivity and ethical issues. Data quality is also an important challenge in which methods for improving data completeness, conformity and plausibility are needed. Further, generalizing and explaining the result of machine learning models are important problems for healthcare, and these are open challenges. A possible solution to improve data quality and accessibility of unstructured data is developing machine learning methods that can generate clinically relevant synthetic data, and accelerating further research on privacy preserving techniques such as deidentification and pseudonymization of clinical text.

This article is categorized under:

Applications of Computational Statistics > Health and Medical Data/Informatics

KEYWORDS

electronic health records, machine learning, statistical methods, unstructured data

1 | INTRODUCTION

The electronic health record (EHR) is defined as a longitudinal collection of electronic health information about individual patients and populations (Kim et al., 2019). EHRs are mainly intended for documentation of the healthcare process. However, EHRs contain a lot of valuable information, which makes them useful for many other purposes. Examples include reduction of medication errors, applying effective methods for communication and sharing information between clinicians, reducing the healthcare costs, better management of patients' medical records, improving the care quality, and leading to better treatment (Kruse et al., 2016). As information about the patient's condition continues to increase rapidly, special expertise for analyzing and extracting information will be required. Data stored in EHR systems can have a variety of formats such as graphics, symbols, free-text, and numbers. These data formats can be classified into structured and unstructured. Examples of structured data include patient demographics (age, gender), height, weight, blood pressure, laboratory tests, and medications. Analysis of structured data types can be performed without much effort using standard statistical or machine learning methods since the data already exists in a fixed structure. Unstructured data, on the other hand, are narrative data like clinical notes, surgical records, discharge summaries, radiology reports, medical images, and pathology reports stored in EHRs (Sun et al., 2018). A lot of valuable information can be extracted from unstructured data, but it is more complicated as they are not in a structured format. For example, unstructured data like free-text discharge summaries contain important information about the care episode or hospital stay, but this information is hard to extract since it is associated with different contexts and contains much uncertainty in medical reporting. In addition, clinical texts include complexities like grammatical and spelling errors, ambiguities, and abbreviations. It should be clear that this type of complexity increases the difficulty of processing and analyzing the data. As the amount of clinical texts increases, methods for analyzing this type of data such as natural language processing (NLP) and deep learning algorithms are gaining wide scholarly interest, especially from data scientists. We describe essential parts of this field in the rest of this paper.

In this review, we will report challenges and opportunities in analyzing both structured and unstructured EHR data. The outline of this paper is as follows. In Section 2, we describe the review methodology and the search criteria. In Section 3, we present a detailed description of unstructured data, the combination of structured and unstructured data, and challenges. Moreover, we focus on several applications of how unstructured data can be utilized in modern healthcare. In Section 4, challenges and bottlenecks of analyzing EHR such as data access, data quality, and privacy preserving methods are discussed. In Section 5, we point out important future research directions before we give our concluding remarks in Section 6.

2 | REVIEW METHODOLOGY

The goal of this review is to provide insights into challenges and opportunities in the analysis of EHRs beyond structured data. We focus on studies that use both structured and unstructured data since the combination is more challenging to analyze in EHR systems. In order to obtain the most relevant challenges and opportunities in the analysis of EHR, we performed our search on two popular databases: PubMed and Web of Science, using the queries as shown in Table 1. Recent, highly relevant, peer-reviewed publications were selected using the following criteria:

- Peer-reviewed publications in journal or conference proceedings.
- Use of machine learning and artificial intelligence (AI) techniques.
- Include methods for analyzing unstructured data.
- Are based on methods which are suitable for heterogeneous data, collected from different sources.
- Can be applied to EHRs.
- Include insight on data quality, data access, and privacy criteria of data.
- Published in recent years (2015–2020).
- Publication language: English.

Database searches (Table 1) identified 423 publications. For providing the background of the subject, 115 research articles in machine learning and unstructured data, but do not include keywords related to challenges and opportunities, were also added. After removing duplicates, 509 publications were considered for further screening. Three coauthors read through the title and abstracts of the papers in the consideration list and eliminated some of the papers.

TABLE 1 Queries used for searching publications

Query no.	Database	Interface	Query (both Boolean search and free text search)	No. of entries
1	PubMed	Advanced	((EHR[Title/Abstract] OR "electronic health record" [Title/Abstract]) OR "electronic health records" [MeSH Terms]) AND (benefits [Title/Abstract] OR challenges [Title/Abstract] OR opportunities [Title/Abstract])) AND "machine learning" [Title/Abstract] AND ("2015/02/09"[PDat]: "2020/02/07"[PDat]	69
2	PubMed	Standard	"Challenges and opportunities of analyzing electronic health records" OR "Challenges of analyzing structured data" OR "Challenges of analyzing unstructured data" OR "Machine learning approach to analysis of structured data" OR "Machine learning approach to analysis of unstructured data" OR "Combination of structured and unstructured data" OR "Challenges of machine learning and artificial intelligence in electronic health records"	160
3	Web of science	Standard	ALL = ((benefits OR challenges OR opportunities) AND (machine learning) AND (EHR OR electronic health record))) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article) Timespan: 2015–2020. Indexes: SCI-EXPANDED, SSCI, AHCI, CPCI-S, CPCI-SSH, ESCI.	194

Further elimination was done based on reading the full-text and 114 publications were regarded as relevant for the review.

3 | UNSTRUCTURED DATA ANALYSIS

In this section, we describe unstructured data and how they can be processed. Unstructured data is data that is not organized in a predefined data model or structure. In this paper, we focus on two main sources of unstructured data in EHR: clinical text and images.

3.1 | Clinical text

Generally, natural language text have historically been avoided by many researchers as well as statisticians since it is difficult to process. Today, more and more methods have emerged to process text using NLP (Jurafsky & Martin, 2019), and text mining (Bird et al., 2009; Dalianis, 2018; Mitkov, 2005).

Natural language text consists of a string of tokens; the tokens are words and delimiters such as space, period, comma, question mark, and so forth. The typical first step an NLP tool performs, before any morphological or linguistic processing, is to tokenize the text, that is, to separate words from delimiters. There are of course difficulties when tokenizing text. For instance, a chunk of text with a measurement of "*0.4 mg*"; should it be kept as one token "20.4 mg," as two tokens "20.4" and "mg," or as four tokens "20," "." "4" and "mg"? The same applies to abbreviations such as "*pat*." Is the period a marker of the end of a sentence or a period in an abbreviation meaning "*pathological*" or "*patient*"?

The special problem with clinical text or electronic patient record text is that it is written in a very specialized language, and there are only a few readily available tools to process it. Second, clinical texts are written in a very telegraphic, information-dense way, for communication between clinicians, and there is a lack of developed dictionaries that can be used for spelling or grammar checking. Also, clinicians sometimes use incomplete sentences, and many times also miss to mention the object such as the patient because the patient is implicit in the text. The patient may be written or just mentioned as "arrived with 38.3 fever and pulse of 132" (Dalianis, 2018).

Natural language text has certain statistical properties. Some words are very common, also known as *stop words*, for example, "*or*," "*if*," "*in*," "*off*," and so forth. However, they do not convey meaningful information. There are around 200 words in each language that are considered as stop words or around 40 percent of the words in a text. The other words belong mainly to one of the classes *noun*, *verbs*, or *adjectives* and these have meaning.

The statistical features of a text are calculated using *term frequency (tf)* and *inverse document frequency (idf)*, also called *tf-idf*. *tf* * *idf* gives a numerical value of how significant a word is in a document collection or corpus. These features are used in information retrieval as well as in cryptography (Van Rijsbergen, 1979).

Natural language text has also syntactical features. The string of words (or tokens) must have a certain order or syntax. To decide on the correct order one needs to use a parser to extract them. A parser is a syntactical analyzer, but usually, it is enough to use a tagger that extracts the words and their syntactic meaning. Each word has also morphological properties with a prefix and or suffix, and sometimes an infix, which can change the meaning of the words. A suffix can indicate whether a noun is in singular or a plural form ("*dog*" or "*dogs*") or the tense of a verb ("*run*," "*ran*," or "*running*"). German and Swedish are compounding languages and the compound can be decompounded to improve the analysis. Knowing these properties of the word makes it possible to normalize the text and slowly transfer it to a structured form so it can be used for machine learning. Features can also be created from the tokens by analyzing the length of tokens, if it is upper or lowercase, first capital letter, numerical, and what token is before or after the analyzed token (Jurafsky & Martin, 2019; Mitkov, 2005).

One part of machine learning is the supervised methods using human annotations. Annotation involves manual labeling or tagging tokens or elements of text with specific identifiers. It means the annotation is carried out to identify important concepts in a text. A tag or an annotation can be for example be set on a symptom, a disorder, a body part, or a drug. Annotation can also be used to classify a whole text to indicate whether a healthcare-associated infection (HAI) has occurred.

These above preprocessing steps to create features are used on textual data before being used by the machine learning algorithms as training data. The evaluation is carried out on unseen test or evaluation data. The performance calculation can be done using metrics such as precision, recall, and F-score (Van Rijsbergen, 1979).

Deep learning techniques take care of all of the above preprocessing challenges by using multiple layers in a neural network to create and represent the features in the textual training data (Goldberg, 2016; Hasan & Farri, 2019). The network is iteratively adjusted with new training data. The data scientist's work is mainly to configure and optimize the algorithms.

3.2 | Clinical images

Although the term "unstructured data" is used more commonly in clinical settings to refer to free text, it also includes several other types of data, such as medical imaging. Medical imaging includes a wide range of categories, from static images produced by diagnostic tools (e.g., radiography) to the videos recorded during medical procedures, such as surgeries. Originally, medical imaging was mainly meant for immediate use, to support a specific diagnostic or therapeutic process for a specific patient. Hence, secondary use of imaging for research was not considered as an option. Thus, medical images and videos were frequently unsaved after their primary use (Kong, 2019). Nowadays, the products of medical imaging seem to have a wider role, being a data source for clinical research, as well as a tool to support care for the individual. As a result, clinical images and video are systematically stored in EHRs, along with the corresponding clinical reports, containing evaluations written by the clinicians who personally visualized and interpreted the image (Assale et al., 2019; Scheurwegs et al., 2016).

Data mining can be performed either on the images themselves or on the corresponding clinical reports (Assale et al., 2019). The latter case falls within the context of clinical text mining, which we discussed extensively in Section 3.1. Data mining on images can be performed with different goals. Some techniques aim to improve the visual quality of clinical imaging, to facilitate its interpretation by a human observer. For example, deep learning methods have been proposed to denoise clinical images Kaji & Kida, 2019; B. Liu & Liu, 2019; Lundervold & Lundervold, 2019). Other data techniques aim to provide clinical decision support. Recently, deep learning techniques have been developed to interpret images in several clinical contexts (Lundervold & Lundervold, 2019). For example, in endoscopy, such techniques can be used for automatic detection and/or classification of lesions based on clinical imaging (Litjens et al., 2017). In dermatology, a common research topic is the automatic classification of skin lesions in terms of malignity; in neurology research effort has been dedicated to the automatic classification of brain diseases (Lundervold & Lundervold, 2019).

To perform image mining, an alternative approach to deep learning is radiomics (Keek et al., 2018; Lambin et al., 2017). According to radiomics, clinical images (e.g., Computerized Tomography, Magnetic Resonance Images, and Positron Emission Tomography) are first converted into a quantitative representation, using a selected set of features. As an example, in oncology, specific features (e.g., shape, dimensions) are extracted from images that represent

cancer. Once the image has been converted into a set of features, it is possible to apply data mining techniques meant for structured data. Commonly, information on the patient's treatment and on its outcome (e.g., survival) is collected along with clinical imaging, and supervised methods can be exploited. This usually leads to a model that can be used to optimize care for a specific patient, given his/her clinical images, in a process known as "precision medicine," or, in this example, "precision oncology."

3.3 | Combining unstructured data with structured data

As stated in the introduction, EHR usually contains both structured and unstructured data. Structured data can be processed relatively easily using various statistical methods. However, structured data alone does not provide all information about the overall clinical context. In contrast, unstructured data can provide extra, valuable information but the analytics processes are complex, time-consuming and requires excessive manual effort. A well-designed diagnosis and decision support tool needs to efficiently use both structured and unstructured data for extracting useful information and provide better outcomes. Several studies have shown the advantages of combining structured and unstructured information. For example, Scheurwegs et al. (2016) presents insights into combination of structured and unstructured data for automating clinical code assignment. They showed that available information in unstructured data is not enough for assigning clinical codes, and that adding structured data improved the performance significantly. Also, Sheikhalishahi et al. (2019) explained how using clinical notes in chronic disease management may assist in early detection of rheumatoid arthritis, Parkinson diseases and Alzheimer disease and help delaying or preventing diseases onset. Furthermore, analysis of unstructured data such as clinical text or images can be enhanced by incorporating the contexts associated with them. For example, diagnosis conducted by clinicians based on clinical images is usually done in combination with structured data such as vital measurement and laboratory tests. These different data sources are complementary with each other and an integration framework of different unstructured and structured data sources is essential to increase the accuracy and reliability of machine learning models. However, as described by Cao et al. (2014), combination of multiple data source can cause redundant and conflicting information.

Combining structured and unstructured data can be done at an early or late stage during the process of data integration (Gligorijević & Pržulj, 2015). Early data integration combines multiple features from different data types before training the model. Model training is then performed on the combined feature set. Late data integration learns a separate model per data source and combines these predictions with a meta-learner (Scheurwegs et al., 2016). The most straightforward method for early data integration is to convert unstructured data to structured data (see Section 3.1). The features of the converted unstructured data will be combined with the features of the structured data sources. A single machine learning model can then be built based on the combined feature vector. This allows unstructured data to be stored and processed similarly to structured. Feature selection can also be done automatically. Cao et al. (2014) proposes a method to automatically select feature by using tensor based multiview method for brain diseases. For example, text could be converted to feature vectors of numerals by using the bag of words model. Then, a prediction model could utilize deep unified networks (Golas et al., 2018). A summary of the studies that combine structured and unstructured data can be found in Table 2.

3.4 | Examples of applications that benefit from using unstructured data

In this section, we describe several applications that use unstructured data to demonstrate how machine learning and statistical tools were implemented. The applications were chosen because diagnoses of conditions related to these applications are highly dependent on unstructured data sources such as clinical text and images.

3.4.1 | Mental health

The large amount of data from EHR has facilitated computational approaches for quality improvement in mental healthcare. For instance, a clinician might want to check the practices of psychiatrists in using metformin to control metabolic dysfunction related to antipsychotics. The simplest cohort for this application can be defined by one variable such as psychosis diagnosis code but dimensional phenotyping is used to extract information from combinations of

TABLE 2 Example of studies using both structured and unstructured data

Reference	Clinical problem	Machine learning method	Data type	Integration strategy	Feature selection
(Golas et al., 2018)	Heart failure	Deep unified networks	Structured EHR and text	Early	Information code mapping, well-known factors mapping, text data selection, and variable coding
(Cao et al., 2014)	Neurological disorder	Support vector machine	Different EHR and images	Early	Dual-Tmfs method
(Scheurwegs et al., 2016)	Assigning clinical codes	Random forests, Naive Bayes and Bayesian networks	Structured EHR and text	Both early and late	Bag of words
(Lei et al., 2020)	Classification of noncommunicable diseases	Deep learning and knowledge graph	Structured EHR, text, image, and video	Early	Correlation analysis
(Shao et al., 2019)	Detection of probable dementia cases	Topic modeling	Structured EHR and text	Early	Correlation and odds ratio analysis
(Zhang et al., 2019)	Electronic phenotyping	Penalized logistic regression	Structured EHR and text	Early	Unsupervised feature learning, ensembled sparse regression
(A. Callahan et al., 2019)	Medical device surveillance	Statistical analysis	Structured EHR and text	Early	Deep learning
(Ross et al., 2019)	Predicting future cardiovascular events	Penalized linear regression and random forest	Structured EHR and text	Early	Annotation
(Murray et al., 2019)	Detecting systemic lupus erythematosus	Logistic regression	Structured EHR and text	Early	Presence of words

structured and unstructured data (Edgcomb & Zima, 2019). In another example, a clinician might want to identify homeless youths that used psychiatric emergency services. Narrative reports from the cohort are processed to transform raw texts into structured data. A combination of narrative and structured data can increase the accuracy of cohort identification and identification of latent cohorts (Edgcomb & Zima, 2019). In another case, a researcher might try to develop an algorithm to predict psychiatric hospital readmission for adolescents with depression and a history of suicide attempts. Narrative discharge summaries with a natural language toolkit are processed to transform narrative terms for suicide attempts to make a new categorical variable in structured data. Machine learning algorithms were used to classify each patient as probable or not probable to be readmitted, using both structured data and the added categorical variable (Edgcomb & Zima, 2019). Several other studies also used NLP to identify suicide ideation and suicide attempts in psychiatric clinical databases and EHR (Downs et al., 2017; Fernandes et al., 2018; Velupillai et al., 2019).

3.4.2 | Detecting and predicting adverse events such as HAI

Healthcare-associated infection is a large problem in healthcare since patients obtain them while being treated and may prolong their healthcare stay and sometimes be lethal. It is estimated that over 10 percent of all inpatients suffer from an HAI. HAIs are underreported but can be detected and sometimes also predicted using EHRs that have been manually annotated for HAI. Each note was manually tagged or annotated as having or not having an HAI by two clinicians jointly. Both support vector machine (SVM) and Gradient Tree Boosting machine learning algorithms were applied and 10-fold cross-validation was used. Gradient Tree Boosting gave the best results (Ehrentraut et al., 2016).

Lots of works have been carried out in HAI detection and prediction. In Freeman et al. (2013), there is a review of different surveillance systems for HAI used worldwide, most of the systems use only the structured information but some of them use both clinical free texts and structured data.

Other adverse effects are adverse drug events (ADEs). In the article by Karimi et al. (2015), the authors review the area of ADE, the type of data where these can be found such as clinical text but also social media, and describes tools using both structured data and text mining techniques to detect ADEs.

3.4.3 | Cancer pathology report coding

Pathology reports that contain information about the stage of a cancer tumor are used both as documentation for treating the patient but also for collecting statistics. For statistical purposes, pathology reports in the Norwegian cancer registry have to be coded. Today, this is a manual and time-consuming process that involves reading the pathology report and then reporting the numbers and data found. This could easily be automated using NLP techniques, specifically machine learning based mining, since there are lots of older manually coded pathology reports (Spasic et al., 2014; Weegar et al., 2017; Weegar & Dalianis, 2015). Weegar and Dalianis (2015) used rule-based algorithms since the pathology reports were well structured.

3.4.4 | Radiology image to radiology report generation

Radiology images and their attached radiology reports can be used to generate automatic radiology reports from unseen radiology images using deep learning, see Figure 1 (Jing et al., 2017).

4 | CHALLENGES AND BOTTLENECKS

In this section, we focus on challenges and bottlenecks related to the analysis of EHR data, in particular with regards to a typical data science pipeline. A broad view of the pipeline could include (i) data access and privacy, (ii) preprocessing, analyses, and model building, and (iii) using the model in production. The pipeline is not necessarily step-wise since the steps can have feedback loops, resulting in a nonsequential process. For instance, more preprocessing and model training may be required after a model is used in production. In succeeding subsections we discuss challenges and bottlenecks commonly associated with the different stages in the pipeline.

4.1 | Data access and privacy

Acquiring the data is an important task because today much research in healthcare data science is "put on hold" due to data inaccessibility. Key infrastructures for secondary use of EHR data are still underdeveloped, and this represents an important research gap.

Access to clinical data such as electronic patient records is very restricted due to the high sensitivity of data accumulated in EHR systems. Ethical permission is needed and also good contact with the healthcare provider. Reuse of such data for research and quality improvement work is regulated at multiple levels (GDPR (2016), HIPAA (2003), and national and institutional guidelines) aiming to ensure the privacy of individuals and healthcare institutions. At the same time, there are forces that push for the use of the valuable information available within the healthcare system to be used not only for research but also for improving healthcare. A proper balance between privacy and utility of personal health data has been studied before, however, no consensus has yet been reached. It still remains a limiting factor in adopting novel data science methodologies in clinical settings.

Several approaches addressing patient privacy concerns when using health data in research have been suggested and successfully applied previously. These techniques can be grouped into two main directions:

a. Obscuring patient identifiers to preserve privacy (deidentification, pseudonymization) and using existing computational techniques on deidentified/pseudonymized data (Du et al., 2018; X. Yang et al., 2019).
b. Adapting existing computational techniques to function in a privacy preserving manner. Improved data privacy comes with costs in terms of algorithmic and infrastructure complexity, reliability of privacy preservation techniques, and utility of obscured data (Giacomelli et al., 2019; Ma et al., 2019).

4.1.1 | Deidentification and pseudonymization

An obvious solution for avoiding disclosure of sensitive information is the removal of patient identifiers (deidentification) or replacing them with meaningful, however, not identifiable data values (pseudonymization). For example, instead of removing names and surnames from clinical texts, these can be replaced with other names and surnames (so-called surrogates or pseudonyms) assigning the same name and surname values to the same individual throughout clinical records. This method keeps the integrity of data while minimizing the reidentification risk.

Regardless of the selected method (deidentification or pseudonymization), finding personal identifiers in clinical data is challenging, especially in unstructured text. However, computational methods show relatively high performance in capturing personal identifiers in clinical notes: Conditional Random Fields model combined with regular expressions achieved F1-score value of 0.9878 in a manually annotated Chinese discharge summary corpus (Du et al., 2018). Similar performance was demonstrated using deep learning methods and English word embeddings in i2b2 corpus (F1-score 0.9584) (X. Yang et al., 2019).

Performance figures of these methods show that computational methods are capable of recognizing sensitive patient details in clinical texts and could be used for deidentification and pseudonymization. However, it is still unclear how well these models generalize and whether the performance is maintained in cross-institutional scenarios.

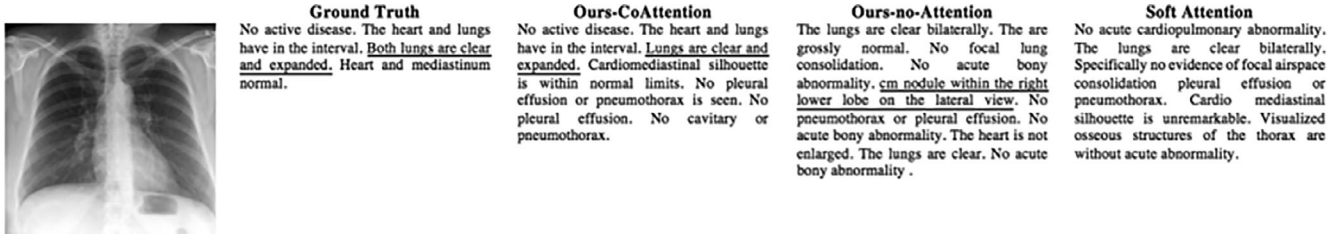

FIGURE 1 Example of a chest X-ray with an attached radiology report written by a radiologist. (Ground truth) and three automatically generated text so-called synthetic texts created by the three different machine learning methods (our-coattention, our-no-attention and soft attention)

Source: Extracted from figure 3 in Jing et al. (2017). Licensed under creative commons

4.1.2 | Adaptation of existing computational techniques to function in a privacy preserving manner

Machine learning based data analytics is typically associated with large data repositories and extensive amounts of computational power. Accumulating health data in single data storage in many cases is not possible due to data privacy concerns, legal regulations, and willingness to store patient data outside health institutions. Cross-border data collection is even more restrictive.

To adapt to the specifics of this strictly regulated landscape, machine learning algorithms are supplemented by privacy preserving features. Accessing data in a federated manner has proven its feasibility and various machine learning algorithms are redeveloped to function in environments where data are distributed across data owners (Giacomelli et al., 2019; Ma et al., 2019). These approaches are built on the idea of performing partial data analysis at each data owner (often in parallel) and aggregating these results at an entire population level. Data owners are able to keep control of their data and have local data access and processing regulations in place, while only aggregated statistics are revealed to the outsiders. Aggregated statistics do not contain sensitive data and patient identifiers and are safe to be shared.

Federated data analytics is a quickly evolving field, however, support for various machine learning algorithms is still limited. While these methods address data privacy and security questions, they introduce additional computational overheads and infrastructural complexity compared to centralized approaches (Giacomelli et al., 2019). Trust in the results produced by federated data analytics infrastructures can also be debated since it is very difficult to validate the accuracy of these results without accumulating data in centralized storage.

4.2 | Preprocessing, analysis, and model building

Two of the most common resources required for preprocessing, analysis, and model building are computing resources and the data itself. In terms of computing resources, some machine learning algorithms require an enormous amount of computing power; both computer memory and processor speed. For example, building language models for natural language understanding (NLU) using deep learning approaches could take weeks on a modern processor or a simple graphics processing unit (GPU). Today, this heavy processing can be done on the cloud using more powerful resources from major cloud services like Google (Hyseni & Ibrahimi, 2017) and Amazon WebServices (AWS) (Ye et al., 2020).

In terms of the data itself, a common challenge, especially for clinical data, is a lack of large-enough labeled data or gold standard corpora. Creating a good corpus is time-consuming and requires specialist domain knowledge. One approach to solving this problem is to use semi-supervised methods, where both supervised and unsupervised methods are used. First, clustering can be used to label data around cluster centers, then the labels are used as additional features for a supervised method. This approach is widely used on healthcare problems, for example, to classify pathology images (Peikari et al., 2018).

Another challenge is the availability of the data in the first place. Some algorithms are data-hungry and this presents challenges for clinical data and for minor languages with fewer resources. There are several solutions to getting around this problem. For NLP problems, one solution is to use natural language generation (NLG) or other artificial methods to generate synthetic text, and then use the text to train NLP models. Synthetic texts have been explored for several tasks in healthcare, including for mental health (Ive et al., 2020) and for extracting adverse events from clinical texts (Tao et al., 2019). However, these solutions can raise questions about data quality.

4.2.1 | EHR data quality challenges

Data quality as described in the included studies can be considered from at least two perspectives. The first relates to data quality assessment (DQA) from the perspective of data acquisition and infrastructure, especially for multi-institutional data grids and systems. A typical DQA framework could involve checking that the data are complete, conformant, and plausible (Kahn et al., 2016; K. Lee et al., 2017). Infrastructure should have well-harmonized data that is mapped onto correct data fields, for example, using common clinical terminology and ontology mapping (T. Callahan et al., 2017; Johnson et al., 2015).

Ontologies and terminologies play a major role in this process. Several classification systems are already used in various domains of healthcare: International Classification of Primary Care (ICPC) codes are used for classifying health problems and diagnosis in primary care, International Statistical Classification of Diseases and Related Health Problems (ICD) codes are used in secondary care, laboratory observations are coded in Logical Observation Identifiers Names and Codes (LOINC). These are only a few out of many coding systems used in healthcare that often have limited interoperability (Ivanović & Budimac, 2014).

Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) is considered the most comprehensive computer processable collection of medical terms. Other medical terminologies, such as ICD-9 and ICD-10 have existing mappings to SNOMED CT making translation between these terminologies consistent. Such mappings have a significant contribution to clinical data quality from a semantic interoperability perspective (El-Sappagh et al., 2018).

At a more general level, interoperable data models ensure that clinical concepts are represented in a consistent manner and have the same semantical meaning across clinical systems. Examples of such clinical data models include openEHR archetypes (Marco-Ruiz et al., 2015) and Fast Healthcare Interoperability Resources (HL7 FHIR) (Saripalle et al., 2019).

Nowadays, HL7 FHIR is one of the most adopted standards to ensure interoperability when exchanging clinical data across different systems. Well-known EHR vendors in several countries are increasingly integrating their software with FHIR. This was also facilitated by joint efforts promoted by HL7 and involving the heterogeneous stakeholders involved in the health care environment. In this regard it is worth mentioning the Argonaut Project (2020), an initiative that established a partnership between leading American EHR vendors (e.g., Epic and Cerner) and major health care providers (e.g., Mayo Clinic and Boston Children's Hospital) to reach a consensus on how to implement systems that adhere to FHIR standards (HL7 Argonaut Project, 2020). The consensus on the implementation requirements is expected to accelerate the adoption of FHIR by further EHR vendors. As an extension of the Argonaut Project, in 2019 HL7 has launched the Accelerator Program (HL7 FHIR Accelerator, 2020; HL7 Vulcan, 2020), an initiative that aims to bridge the gaps between health care and research, facilitating the secondary use of the data collected in clinical practice, leveraging on FHIR. To reach that goal, the Accelerator Program runs several projects, that establish collaborations between the different stakeholders, including EHR vendors, research centers, government agencies, patient organizations, health care centers, and organizations for standards development. The projects focus on different use cases, according to the considered clinical domain. For example, in oncology the project, CodeX (Common Oncology Data Elements eXtensions) (HL7 Codex, 2020) aims to facilitate the reuse of EHR data related to cancer treatment for clinical trials, preserving the patient's privacy.

According to the described initiatives, it seems that the different stakeholders are aware of the need to bridge the gap between clinical practice and research, and they agree that standards represent an indispensable tool for achieving that goal.

The other perspective considers data quality at the analysis stage. At this stage, several techniques are used to deal with data quality issues such as sparse data and imputing missing values (Y. Liu & Gopalakrishnan, 2017), detecting outliers (Estiri et al., 2019) and sampling imbalanced data with skewed distributions. Studies taking this perspective considered both structured and unstructured data.

EHR records will contain implausible values because of human error, and therefore detecting outliers is important to ascertain the extent of potential problems that may render any analysis on the data invalid. Traditional abnormal detection systems use distance measures or standard deviation, but one study reported that clustering approaches yielded superior results (Estiri et al., 2019).

Another important challenge is dealing with missing and sparse EHR data, whether the mechanism is missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (Bhaskaran & Smeeth, 2014). The simplest solution could be to omit incomplete records from further analysis, but this may introduce further problems. In addition to the risk of introducing bias, omitting records may result in a very small sample size.

To counter the negative effects of missing data, multiple imputations is a commonly used approach. Both unsupervised (e.g., clustering with k-nearest neighbors-kNN, self-organizing maps), supervised (e.g., decision trees), and more recently, deep learning (Beaulieu-Jones & Moore, 2017), have been explored to impute missing data. Some studies have reported little to no effect of data imputation on the performance of the final model (Y. Liu & Gopalakrishnan, 2017), while other studies have suggested that the nature of the data will dictate the most appropriate imputation method. The effect of these data imputation methods on the final model performance is still not well-understood.

4.2.2 | Challenges analyzing free-text clinical notes in EHR

From the studies we considered, one of the most popular application areas for NLP in healthcare is adverse events detection. The goal with this use-case is to use additional information from texts to support current reporting and alert mechanisms for adverse events and incidents reporting related to patient safety. This also includes pharmacovigilance, ADEs, and general medication safety.

Examples of challenges related to common use-cases include determining temporality from clinical text to ascertain the order of the events, such as point of symptom onset (Juhn & Liu, 2020). Another challenge is determining the source of the adverse event because determining coreference within text or determining relations among entities or clinical concepts in clinical texts is an open challenge (Sheikhalishahi et al., 2019).

One of the most cited challenges is the deidentification of free-text clinical notes. Current deidentification systems work better on structured data than on free text, and deidentification of clinical texts remains an open challenge. More than just detecting identifying personal information in text, the data may contain logical loop-holes that may make it possible to reidentify a person. Using additional data sources that are independent of the clinical text, it may be possible to combine the information to identify persons; due caution is required to assess reidentification risk (Simon et al., 2019). Therefore, even though clinical text can be said to be deidentified, it is still prudent to use the text responsibly.

While combining data sources can pose security risks, it can be beneficial to get a complete view of the record, for analysis purposes. There is often a lack of information exchange between data sources, for example within and without hospital systems. Clinical data can reside in organizational silos because the different systems are not able to communicate, and this poses interoperability challenges. A related challenge to interoperability is information heterogeneity, for example, when analyzing text that was written by clinicians with different writing and organizational styles.

4.3 | Using the model in production

While the performance of AI systems has significantly improved from the early days and has also become more pervasive in society, there are still significant challenges when implementing them in healthcare institutions. However, in this section, we focus on the technological aspects, rather than the human factors or organizational challenges.

4.3.1 | Explainability and biases of the models

Model interpretability or explainability (Ribeiro et al., 2016) includes getting insight into how the model generated the results. In healthcare, black-box solutions can lead to problems with technology adoption since clinicians need to understand and explain the decisions for which they are ultimately responsible. While the new GDPR encourages AI systems to explain their results, explainable results are not set as a legal requirement (Wachter et al., 2018).

Some machine learning models lend themselves well to interpretable and explainable results, while others do not. For example, it is easy to inspect how a decision tree arrives at the results. This is not the case with some recent neural network-based models, with deep hidden layers, where it is not possible to readily explain how the model arrived at the result.

In healthcare, being able to explain the results is crucial for clinicians and other healthcare professionals, before they can place their trust in the results from the models. Theoretically, a complex neural network model can discover novel patterns that are beyond human comprehension and requiring these models to be interpretable or explainable places a constraint on the new knowledge these models might discover. Nonetheless, explainable models in healthcare is a topic that is gaining wide scholarly interest (Choo & Liu, 2018; H. Lee et al., 2019).

Another challenge that can become more apparent while using the model in production is the bias, therefore it is important to ascertain sources of bias during the model training phases. One possible source of bias stems from selection bias. Training data from clinical sources is usually from specific healthcare units and the patients do not represent a random sample from the whole population.

4.3.2 | Generalizability, transferability, and updating the models

Model generalizability or transferability refers to the ability to adapt models trained on a specific dataset for use on other datasets, say, at a different hospital department or on an entirely different health institution. This is considered one of the most important challenges for the actual use of AI at health institutions (Bates et al., 2020). The current status is that some degree of fine-tuning or optimization is required for the model to perform comparatively as well on the new dataset.

Models based on rule-based methods can sometimes be more difficult to generalize because they require a lot of manual effort to build. A slight variation in the data may require even more effort to construct new rules. Machine learning based methods are more robust and are able to learn patterns that enable them to be more easily adaptable to variations in data. However, they still normally require additional training on the new data.

A promising solution for generalizing AI models is to use transfer learning (Q. Yang et al., 2020). For example, a model built using data from the general domain can be fine-tuned or adapted to perform nearly as well in the clinical domain, using transfer learning techniques. The fine-tuning or "transfer" process requires a much smaller training dataset than would normally be required. Transfer learning using contextual embeddings (Devlin et al., 2018) that are specially trained on health data (J. Lee et al., 2020) is gaining wide attention for solving problems in healthcare such as extracting clinical concepts from clinical text (Si et al., 2019).

5 | FUTURE RESEARCH DIRECTIONS

As pointed out above, there have been several attempts to extract important information from EHR data using machine learning and statistical methods (Khalifa, 2019; Nair et al., 2016; Ratwani et al., 2016; Wang & Preininger, 2019; Xiao et al., 2018). From these attempts, it is clear that such methods may lead to useful decision support tools and recommendation systems for both patients and healthcare personnel. In the future, such tools will very likely be an integrated part of the care general practitioners (GPs) and hospitals can offer (Nair et al., 2016). Research in a large number of directions is, however, needed before such systems can eventually be put into use. Some of these research questions can be solved on a reasonably short-horizon while others are far into the future. Below, we start out by pointing out some specific short-term goals where researchers in machine learning and statistical methodology are expected to have important impacts. Next, we describe long-term goals that can be obtained by using all available information in the search for continuously improved healthcare solutions.

5.1 | Short-term goals

An immediate direction for statisticians lies in the development of systems that contain reliable uncertainty estimation. For applications within healthcare, this is of particular interest, since the users need to have a clear idea about the uncertainty of a suggested decision. Bayesian methods are a natural starting point for this type of research (Broekhuizen et al., 2015; Jothi et al., 2015; Langarizadeh & Moghbeli, 2016).

In Section 4, we underlined that privacy issues represent a severe bottleneck in the analysis of sensitive information. Because of this, it is natural to look for alternative solutions. A natural and obvious alternative would be to analyze a population of "similar patients" (Parimbelli et al., 2018; Sharafoddini et al., 2017). Using this approach, we work on the distributional aspects of a patient's condition so privacy concerns will not be an issue (Hailemichael et al., 2015). Developed methods based on data from the whole population can next be used to suggest successful treatments for future patients at, for example, the GPs' office (Kueper et al., 2020). Such methods will not be as precise as a targeted treatment for a patient's condition, but it is still expected to be very useful.

An alternative procedure would be to utilize tools such as SynthPop (Nowok et al., 2016) and Synthea (Walonoski et al., 2018) to create synthetic data. Synthetic data is data that is artificially created or generated by algorithms, rather than actual measurements. An important task here would be to analyze whether such a dataset has the needed resemblance with real data (Reiner Benaim et al., 2020). Research in this direction would make it possible to obtain realistic synthetic data sets (Chen et al., 2019). If this is possible, privacy issues could be bypassed and a boost in statistical methodology utilizing EHR data could be obtained. Completely realistic data are of course very hard, if not impossible, to obtain, but research in this direction is still very important because it would make a large number of such data sets

available and thereby increase the contribution of statistical methods significantly in this important, but challenging field.

As mentioned in previous paragraphs, sensitive information must be handled with care to avoid misuse (GDPR, 2016). This is frequently mentioned as a bottleneck in the development of new medical systems, but little information is given about the likelihood of such misuse (S. Yu, 2016). Research in this direction would shed light on this important problem, and it might give indications about how strict an acceptable system must be. In almost any application there are unwanted incidents, but they do still happen. In the future, it may, therefore, be necessary to discuss whether a system that handles privacy issues with a specified probability of privacy loss close to zero is sufficient.

Nowadays, there is a rapid increase in AI solutions offered within healthcare (D. Miller & Brown, 2018; Tekkeşin, 2019; Wang & Preininger, 2019; K.-H. Yu et al., 2018). A thorough evaluation of such systems' success is needed to be able to keep the successful ones and get rid of the others. Statistical methods can play an important role in the evaluation of such systems. Moreover, they can be used to justify in what way a decision support tool improves a system.

Research on unstructured data alone is a somewhat new field for many statisticians, but we believe that researchers with a background in machine learning and statistics can make a difference in this area. The underlying idea is of course to extract information from all kinds of unstructured data, but it makes sense to start out with each source separately. Extracting important knowledge from text like nurse notes is an important example of such data. Preliminary results (see, e.g., Bjarnadottir & Lucero, 2018; Korach et al., 2020; Soguero-Ruiz et al., 2016; Topaz et al., 2016; Topaz et al., 2019), indicate that this information typically adds incremental value. The information is, however, hidden in texts that are really hard to interpret (Dalianis, 2018). Even simple methods like "bag-of-words," which essentially describes the frequency of words in the notes, can be useful. Since the condition of a patient can be expressed in many different ways as unstructured texts, it is of vital importance for researchers in machine learning and statistics to utilize the power of NLP in their work, (note that for each language one different NLP system is needed [Névéol et al., 2018]). This is an area where people in the statistical community can make important contributions in the future.

5.2 | Long-term goals

Images represent another type of unstructured data and a really exciting and useful application would be to detect changes in images over time. This could be useful in several areas, in particular for screenings of breast and colon cancer. Nowadays, the detection of harmful changes is left to physicians. This is a very challenging task even for trained experts and in addition, it is very time-consuming. The development of successful decision support tools in this area would, therefore, be extremely useful for the patients, physicians, and society at large. To our knowledge, no successful system of this type exist, although preliminary methods to apply temporal data mining on clinical images have been proposed to monitor the progress of other diseases, such as lung cancer (Singh et al., 2018), and glioblastoma (Smedley et al., 2018). In particular, in Singh et al. (2018), the authors propose a deep learning method to monitor the presence of abnormalities in chest radiographs over time. In Smedley et al. (2018), magnetic resonance images over time and information on the patient's treatment are used to produce logistic regression models for predicting survival in patients affected by glioblastoma. The lack of literature on applications of temporal data mining on images indicates that this is a really challenging problem, but the benefits of a successful methodology make these efforts well worthwhile.

Developing systems for highly heterogeneous data sets is another interesting area. In particular, novel methods that can combine unstructured and structured data are expected to be extremely important in future innovations. A natural starting point here would be to suggest such methods and thereafter illustrate that the combined information improves the outcome in terms of, for example, early detection of cancer or other harmful conditions.

For many groups of patients, it will also be crucial to have methods that can extract information from EHR and self-gathered data. As an example, chronic pain patients constitute a large group of people that frequently are neglected in the health system due to a lack of useful treatments (Crofford, 2015; Mills et al., 2016). The more systematic use of EHR and self-reported data may in the future lead to systems that will increase their quality of life (Kooij et al., 2017; Milani & Franklin, 2017).

Patients with multimorbidity are another group of patients where small improvements in treatment may lead to boosts for patients and society (Muth et al., 2019). The underlying idea here is to utilize all available information and offer personalized treatment. Information about similar patients is an important ingredient also here. Designing successful decision support tools for this group is a challenging task where researchers within machine learning and

statistical methodology can contribute. By using information from similar patients, it is possible to utilize the patient's needs and wishes and thereby offer a safe, personalized treatment for a very challenging group of patients that frequently have a poor quality of life and costs society large amounts of money every year. Successful systems will be able to improve their quality of life at the same time as societal costs are drastically reduced.

At some point in the future, decision support tools will be an integrated part of all areas of healthcare (Khong et al., 2015; Sittig et al., 2016; Wasylewicz & Scheepers-Hoeks, 2018). At the GP office, such systems can be used as a second opinion after the GP has arrived at his initial conclusion utilizing his knowledge and all existing information. This may be helpful in the sense that it can see patterns in a complicated data set that is not visible to the GP. Clearly, it can also lead to situations where the system in fact is misleading the GP. Because of this, such systems will have to be used with caution and they should learn from mistakes so that the given suggestions improve along time (Khairat et al., 2018; K. Miller et al., 2017).

In hospitals, decision support tools learning over time will presumably play an even more important role (Gardner, 2016; White et al., 2017). Such systems will be important both in routine controls and for suggesting the best actions in situations with an overload of information. A change in this direction within hospitals will also mean that more people with expertise in machine learning and statistical methodology must be an integrated part of the hospital staff. Interdisciplinary teams will very likely discover patterns that would otherwise not have been revealed and this can be useful both during regular care, intensive care and to prevent future harmful conditions. Successful contributions within information retrieval from all available data will pave the way for improved healthcare solutions in the future. We believe that strong interdisciplinary teams, including clinicians and researchers within machine learning and statistical methodology, will be the drivers of this development.

6 | CONCLUSION

Unstructured data has a large potential to provide valuable information for health analytics. However, there are still many open research questions that must be addressed before unstructured data can be effectively used in decision support tools and recommendation systems for both patients and healthcare personnel:

- How to simplify the process of unstructured data analysis and reduce manual effort?
- How to protect the privacy and security, improve data quality and accessibility of unstructured data?

To answer this, obscuring patient identifiers to preserve privacy (deidentification, pseudonymization) or synthetic data generation are directions that need to be explored further. Novel methods that can combine unstructured and structured data are expected to be extremely important in future innovations. For many groups of patients, it will also be crucial to have methods for extracting information from EHR and self-gathered data, to improve data accessibility.

ACKNOWLEDGMENTS

Tromsø Forskningsstiftelse, project title: "A smart controller for T1D using RL and SS representation," Grant/award number: A3327. Northern Norway Regional Health Authority, project title: "NorKlinTekst: Natural language processing to extract knowledge from clinical notes in electronic health records," Grant/award number: HNF1395-18.

AUTHOR CONTRIBUTIONS

Maryam Tayefi: Writing-original draft; writing-review and editing. **Phuong Ngo:** Writing-original draft; writing-review and editing. **Taridzo Chomutare:** Writing-original draft; writing-review and editing. **Hercules Dalianis:** Writing-original draft; writing-review and editing. **Elisa Salvi:** Writing-original draft; writing-review and editing. **Andrius Budrionis:** Writing-original draft; writing-review and editing. **Fred Godtliebsen:** Writing-original draft; writing-review and editing.

ORCID

Fred Godtliebsen https://orcid.org/0000-0001-7896-8634

RELATED WIRES ARTICLES

Recent advances in hyperspectral imaging for melanoma detection

REFERENCES

Assale, M., Dui, L., Cina, A., Seveso, A., & Cabitza, F. (2019). The revival of the notes field: Leveraging the unstructured content in electronic health records. *Frontiers of Medicine (Lausanne)*, *6*, 66.

Bates, D. W., Auerbach, A., Schulam, P., Wright, A., & Saria, S. (2020). Reporting and implementing interventions involving machine learning and artificial intelligence. *Annals of Internal Medicine*, *172*(11_Supplement), S137–S144. https://doi.org/10.7326/m19-0872

Beaulieu-Jones, B., & Moore, J. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. *Pacific Symposium on Biocomputing*, *22*, 207–218.

Bhaskaran, K., & Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, *43*, 1336–1339.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Bjarnadottir, R., & Lucero, R. (2018). What can we learn about fall risk Factors from EHR nursing notes? A text mining study. *eGEMs (Washington, DC)*, *6*, 21.

Broekhuizen, H., Groothuis-Oudshoorn, C. G. M., van Til, J. A., Hummel, J. M., & IJzerman, M. J. (2015). A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions. *PharmacoEconomics*, *33*(5), 445–455. https://doi.org/10.1007/s40273-014-0251-x

Callahan, T. J., Bauck, A. E., Bertoch, D., Brown, J., Khare, R., Ryan, P. B., ... Kahn, M. G. (2017). A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, *5*(1), 8–8. http://dx.doi.org/10.5334/egems.223.

Callahan, A., Fries, J. A., Ré, C., Huddleston, J. I., Giori, N. J., Delp, S., & Shah, N. H. (2019). Medical device surveillance with electronic health records. *npj Digital Medicine*, *2*(1), 94. https://doi.org/10.1038/s41746-019-0168-z

Cao, B., He, L., Kong, X., Yu, P., Hao, Z., & Ragin, A. (2014, December). Tensor-based multi-view feature selection with applications to brain diseases. *Proceedings of the IEEE international conference on data mining*, 2014, 40–49.

Chen, J., Chun, D., Patel, M., Chiang, E., & James, J. (2019). The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Medical Informatics and Decision Making*, *19*, 44.

Choo, J., & Liu, S. (2018). Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, *38*, 84–92.

Crofford, L. (2015). Chronic pain: Where the body meets the brain. *Transactions of the American Clinical and Climatological Association*, *126*, 167–183.

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer. https://doi.org/10.1007/978-3-319-78503-5

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Downs, J., Velupillai, S., George, G., Holden, R., Kikoler, M., Dean, H., ... Dutta, R. (2017). Detection of suicidality in adolescents with autism spectrum disorders: Developing a natural language processing approach for use in electronic health records. *American Medical Informatics Association annual symposium proceedings*, 2017, 641–649.

Du, L., Xia, C., Deng, Z., Lu, G., Xia, S., & Ma, J. (2018). A machine learning based approach to identify protected health information in Chinese clinical text. *International Journal of Medical Informatics*, *116*, 24–32. https://doi.org/10.1016/j.ijmedinf.2018.05.010

Edgcomb, J., & Zima, B. (2019). Machine learning, natural language processing, and the electronic health record: Innovations in mental health services research. *Psychiatric Services*, *70*, 346–349.

Ehrentraut, C., Ekholm, M., Tanushi, H., Tiedemann, J., & Dalianis, H. (2016). Detecting hospital-acquired infections: A document classification approach using support vector machines and gradient tree boosting. *Health Informatics Journal*, *24*, 24–42.

El-Sappagh, S., Franda, F., Ali, F., & Kwak, K. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making*, *18*, 76.

Estiri, H., Klann, J. G., & Murphy, S. N. (2019). A clustering approach for detecting implausible observation values in electronic health records data. *BMC Medical Informatics and Decision Making*, *19*(1), 142. https://doi.org/10.1186/s12911-019-0852-6

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific Reports*, *8*(1), 7426. https://doi.org/10.1038/s41598-018-25773-2

Freeman, R., Moore, L., Garcia, A. L., Charlett, A., & Holmes, A. (2013). Advances in electronic surveillance for healthcare-associated infections in the 21st century: A systematic review. *The Journal of Hospital Infection*, *84*, 106–119.

Gardner R. M. (2016). Clinical Information Systems – From Yesterday to Tomorrow. *Yearbook of Medical Informatics*, *25*, (S 01), S62–S75. http://dx.doi.org/10.15265/iys-2016-s010.

GDPR. (2016). *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data*. EU, 679. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Giacomelli, I., Jha, S., Kleiman, R., Page, D., & Yoon, K. (2019). Privacy-preserving collaborative prediction using random forests. *AMIA joint summits on translational science proceedings*, 2019, 248–257.

Gligorijević, V., & Pržulj, N. (2015). Methods for biological data integration: Perspectives and challenges. *Journal of The Royal Society Interface*, *12*(112), 20150571. https://doi.org/10.1098/rsif.2015.0571

Golas S. B., Shibahara T., Agboola S., Otaki H., Sato J., Nakae T., Hisamitsu T., Kojima G., Felsted J., Kakarmath S., Kvedar J., Jethwani K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of

electronic medical records data. *BMC Medical Informatics and Decision Making*, *18*, (1), 1–44. http://dx.doi.org/10.1186/s12911-018-0620-z.

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*, 345–420. https://doi.org/10.1613/jair.4992 https://jair.org/index.php/jair/article/view/11030

Hailemichael, M., Marco-Ruiz, L., & Bellika, J. (2015). Privacy-preserving statistical query and processing on distributed OpenEHR data. *Studies in Health Technology and Informatics*, *210*, 766–770.

Hasan, S. A., & Farri, O. (2019). Clinical natural language processing with deep learning. In *Data science for healthcare* (pp. 147–171). Springer International Publishing. https://doi.org/10.1007/978-3-030-05249-2_5 http://link.springer.com/10.1007/978-3-030-05249-2{_}5

HIPAA. (2003). *Health insurance portability and accountability act (HIPAA). U.S. Department of Health and Human Services*. http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm

HL7 Argonaut Project. (2020). *HL7 Argonaut Project*. https://argonautwiki.hl7.org/Main_Page

HL7 Codex. (2020). *HL7 Codex*. https://www.hl7.org/codex/

HL7 FHIR Accelerator. (2020). *HL7 FHIR Accelerator*. https://www.hl7.org/about/fhir-accelerator/

HL7 Vulcan. (2020). *HL7 Vulcan*. http://www.hl7.org/vulcan/

Hyseni, L. N., & Ibrahimi, A. (2017). Comparison of the cloud computing platforms provided by Amazon and Google. In *2017 Computing Conference* (pp. 236–243). IEEE. https://doi.org/10.1109/SAI.2017.8252109 http://ieeexplore.ieee.org/document/8252109/

Ivanović, M., & Budimac, Z. (2014). An overview of ontologies and data resources in medical domains. *Expert Systems with Applications*, *41* (11), 5158–5166. https://doi.org/10.1016/j.eswa.2014.02.045

Ive, J., Viani, N., Kam, J., Yin, L., Verma, S., Puntis, S., Cardinal, R. N., Roberts, A., Stewart, R., & Velupillai, S. (2020). Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, *3*(1), 69. https://doi.org/10.1038/s41746-020-0267-x

Jing, B., Xie, P., & Xing, E. (2017). *On the automatic generation of medical imaging reports*. arxiv preprint arxiv:1711.08195.

Johnson, S., Speedie, S., Simon, G., Kumar, V., & Westra, B. (2015). A data quality ontology for the secondary use of EHR data. In *American Medical Informatics Association annual symposium proceedings, 2015*, 1937–1946.

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare—A review. *Procedia Computer Science*, *72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Juhn, Y., & Liu, H. (2020). Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *The Journal of Allergy and Clinical Immunology*, *145*, 463–469.

Jurafsky, D., & Martin, J. H. (2019). *Speech and language processing*. Pearson https://web.stanford.edu/~jurafsky/slp3/

Kahn, M., Callahan, T., Barnard, J., Bauck, A., Brown, J., Davidson, B., … Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs (Washington, DC)*, *4*, 1244.

Kaji, S., & Kida, S. (2019). Overview of image-to-image translation by use of deep neural networks: Denoising, super-resolution, modality conversion, and reconstruction in medical imaging. *Radiological Physics and Technology*, *12*, 235–248.

Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, E., & Paris, C. (2015). Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, *47*(4), 56.

Keek, S. A., Leijenaar, R. T., Jochems, A., & Woodruff, H. C. (2018). A review on radiomics and the future of theranostics for patient selection in precision medicine. *The British Journal of Radiology*, *91*(1091), 20170926. https://doi.org/10.1259/bjr.20170926

Khairat, S., Marc, D., Crosby, W., & Sanousi, A. A. (2018). Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Medical Informatics*, *6*(2), e24. https://doi.org/10.2196/medinform.8912

Khalifa, M. (2019). Challenges of health analytics utilization: A review of literature. *Studies in Health Technology and Informatics*, *262*, 55–58.

Khong, P., Holroyd, E., & Wang, W. (2015). A critical review of the theoretical frameworks and the conceptual factors in the adoption of clinical decision support systems. *Computers, Informatics, Nursing*, *33*, 555–570.

Kim, E., Rubinstein, S., Nead, K., Wojcieszynski, A., Gabriel, P., & Warner, J. (2019). The evolving use of electronic health records (EHR) for research. *Seminars in Radiation Oncology*, *29*, 354–361.

Kong, H. (2019). Managing unstructured big data in healthcare system. *Healthcare Informatics Research*, *25*, 1–2.

Kooij, L., Groen, W., & van Harten, W. (2017). The effectiveness of information technology-supported shared care for patients with chronic disease: A systematic review. *Journal of Medical Internet Research*, *19*, e221.

Korach, Z., Yang, J., Rossetti, S., Cato, K., Kang, M., Knaplund, C., … Zhou, L. (2020). Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. *International Journal of Medical Informatics*, *135*, 104053.

Kruse, C. S., Kristof, C., Jones, B., Mitchell, E., & Martinez, A. (2016). Barriers to electronic health record adoption: A systematic literature review. *Journal of Medical Systems*, *40*(12), 252. https://doi.org/10.1007/s10916-016-0628-9

Kueper, J., Terry, A., Zwarenstein, M., & Lizotte, D. (2020). Artificial intelligence and primary care research: A scoping review. *Annals of Family Medicine*, *18*, 250–258.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., de Jong, E. E., van Timmeren, J., … Walsh, S. (2017). Radiomics: The bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, *14*(12), 749–762. https://doi.org/10.1038/nrclinonc.2017.141

Langarizadeh, M., & Moghbeli, F. (2016). Applying naive Bayesian networks to disease prediction: A systematic review. *Acta Informatica Medica*, *24*(5), 364–369. https://doi.org/10.5455/aim.2016.24.364-369

Lee, K., Weiskopf, N., & Pathak, J. (2017). A framework for data quality assessment in clinical research datasets. *American Medical Informatics Association annual symposium proceedings, 2017*, 1080–1089.

Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S., Guerrier, C., ... Do, S. (2019). An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, *3*, 173–182.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*, 1234–1240.

Lei, Z., Sun, Y., Nanehkaran, Y., Yang, S., Islam, M. S., Lei, H., & Zhang, D. (2020). A novel data-driven robust framework based on machine learning and knowledge graph for disease classification. *Future Generation Computer Systems*, *102*, 534–548. https://doi.org/10.1016/j.future.2019.08.030

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A. W. M., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu Yuzhe, Gopalakrishnan Vanathi (2017). An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data. *Data*, *2*, (1), 8http://dx.doi.org/10.3390/data2010008.

Liu, B., & Liu, J. (2019). Overview of image denoising based on deep learning. *Journal of Physics: Conference Series*, *1176*, 022010. https://doi.org/10.1088/1742-6596/1176/2/022010

Lundervold, A., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, *29*, 102–127.

Ma, J., Zhang, Q., Lou, J., Ho, J., Xiong, L., & Jiang, X. (2019). Privacy-preserving tensor factorization for collaborative health data analysis. In *Proceedings of the ACM international conference on information and knowledge management, 2019*, 1291–1300.

Marco-Ruiz, L., Moner, D., Maldonado, J., Kolstrup, N., & Bellika, J. (2015). Archetype-based data warehouse environment to enable the reuse of electronic health record data. *International Journal of Medical Informatics*, *84*, 702–714.

Milani, R. V., & Franklin, N. C. (2017). The role of technology in healthy living medicine. *Progress in Cardiovascular Diseases*, *59*(5), 487–491. https://doi.org/10.1016/j.pcad.2017.02.001

Miller, D., & Brown, E. (2018). Artificial intelligence in medical practice: The question to the answer? *The American Journal of Medicine*, *131*, 129–133.

Miller, K., Mosby, D., Capan, M., Kowalski, R., Ratwani, R., Noaiseh, Y., ... Arnold, R. (2017). Interface information, interaction: A narrative review of design and functional requirements for clinical decision support. *Journal of the American Medical Informatics Association*, *25*(5), 585–592. https://doi.org/10.1093/jamia/ocx118

Mills, S., Torrance, N., & Smith, B. H. (2016). Identification and management of chronic pain in primary care: A review. *Current Psychiatry Reports*, *18*(2), 22. https://doi.org/10.1007/s11920-015-0659-9

Mitkov, R. (2005). *The Oxford handbook of computational linguistics*. Oxford University Press.

Murray, S., Avati, A., Schmajuk, G., & Yazdany, J. (2019). Automated and flexible identification of complex disease: Building a model for systemic lupus erythematosus using noisy labeling. *Journal of the American Medical Informatics Association*, *26*, 61–65.

Muth C., Blom J. W., Smith S. M., Johnell K., Gonzalez-Gonzalez A. I., Nguyen T. S., Brueckle M.-S., Cesari M., Tinetti M. E., Valderas J. M. (2019). Evidence supporting the best clinical management of patients with multimorbidity and polypharmacy: a systematic guideline review and expert consensus. *Journal of Internal Medicine*, *285*(3), 272–288. http://dx.doi.org/10.1111/joim.12842.

Nair, S., Hsu, D., & Celi, L. (2016). *Challenges and opportunities in secondary analyses of electronic health record data*.

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: Opportunities and challenges. *Journal of Biomedical Semantics*, *9*, 12.

Nowok B., Raab G. M., Dibben C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74*, (11), 1–26. http://dx.doi.org/10.18637/jss.v074.i11.

Parimbelli, E., Marini, S., Sacchi, L., & Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*, *83*, 87–96.

Peikari, M., Salama, S., Nofech-Mozes, S., & Martel, A. (2018). A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific Reports*, *8*, 7193.

Ratwani, R., Fairbanks, T., Savage, E., Adams, K., Wittie, M., Boone, E., ... Gettinger, A. (2016). Mind the gap. A systematic review to identify usability and safety challenges and practices during electronic health record implementation. *Appl Clin Inform*, *7*, 1069–1087.

Reiner Benaim, A., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., Khamaisi, M., Lurie, Y., Azzam, Z. S., Khoury, J., Kurnik, D., & Beyar, R. (2020). Analyzing medical research results based on synthetic data and their relation to real data results: Systematic comparison from five observational studies. *JMIR Medical Informatics*, *8*(2), e16492. https://doi.org/10.2196/16492 http://medinform.jmir.org/2020/2/e16492/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should I trust you? In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM. https://doi.org/10.1145/2939672.2939778

Ross, E., Jung, K., Dudley, J., Li, L., Leeper, N., & Shah, N. (2019). Predicting future cardiovascular events in patients with peripheral artery disease using electronic health record data. *Circulation. Cardiovascular Quality and Outcomes*, *12*, e004741.

Saripalle, R., Runyan, C., & Russell, M. (2019). Using HL7 FHIR to achieve interoperability in patient health record. *Journal of Biomedical Informatics*, *94*, 103188.

Scheurwegs, E., Luyckx, K., Luyten, L., Daelemans, W., & Van, d. B. T. (2016). Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *Journal of the American Medical Informatics Association*, *23*, e11–e19.

Shao, Y., Zeng, Q., Chen, K., Shutes-David, A., Thielke, S., & Tsuang, D. (2019). Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Medical Informatics and Decision Making*, *19*, 128.

Sharafoddini, A., Dubin, J., & Lee, J. (2017). Patient similarity in prediction models based on health data: A scoping review. *JMIR Medical Informatics*, *5*, e7.

Sheikhalishahi, S., Miotto, R., Dudley, J., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, *7*, e12239.

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, *26*, 1297–1304.

Simon, G., Shortreed, S., Coley, R., Penfold, R., Rossom, R., Waitzfelder, B., ... Lynch, F. (2019). Assessing and minimizing re-identification risk in research data derived from health care records. *eGEMs (Washington, DC)*, *7*, 6.

Singh, R., Kalra, M. K., Nitiwarangkul, C., Patti, J. A., Homayounieh, F., Padole, A., Rao, P., Putha, P., Muse, V. V., Sharma, A., & Digumarthy, S. R. (2018). Deep learning in chest radiography: Detection of findings and presence of change. *PLoS One*, *13*(10), e0204155. https://doi.org/10.1371/journal.pone.0204155

Sittig, D. F., Wright, A., & Middleton, B. (2016). Clinical decision support: A 25 year retrospective and a 25 year vision. *Yearbook of Medical Informatics*, *25*(S 01), S103–S116. https://doi.org/10.15265/iys-2016-s034

Smedley, N. F., Ellingson, B. M., Cloughesy, T. F., & Hsu, W. (2018). Longitudinal patterns in clinical and imaging measurements predict residual survival in glioblastoma patients. *Scientific Reports*, *8*(1), 14429. https://doi.org/10.1038/s41598-018-32397-z

Soguero-Ruiz, C., Hindberg, K., Rojo-Alvarez, J., Skrovseth, S., Godtliebsen, F., Mortensen, K., ... Jenssen, R. (2016). Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. *IEEE Journal of Biomedical and Health Informatics*, *20*, 1404–1415.

Spasic, I., Livsey, J., Keane, J., & Nenadic, G. (2014). Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics*, *83*(9), 605–623.

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: A review. *Journal of Healthcare Engineering*, *2018*, 1–9. https://doi.org/10.1155/2018/4302425

Tao, C., Lee, K., Filannino, M., & Uzuner, O. (2019). An exploratory study on pseudo-data generation in prescription and adverse drug reaction extraction. *Studies in Health Technology and Informatics*, *264*, 388–392.

Tekkeşin, A. (2019). Artificial intelligence in healthcare: Past, present and future. *Anatolian Journal of Cardiology*, *22*, 8–9.

Topaz, M., Radhakrishnan, K., Lei, V., & Zhou, L. (2016). Mining clinicians' electronic documentation to identify heart failure patients with ineffective self-management: A pilot text-mining study. *Studies in Health Technology and Informatics*, *225*, 856–857.

Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., & Bowles, K. H. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, *90*, 103103. https://doi.org/10.1016/j.jbi.2019.103103

Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworth & Co.

Velupillai, S., Epstein, S., Bittar, A., Stephenson, T., Dutta, R., & Downs, J. (2019). Identifying suicidal adolescents from mental health records using natural language processing. *Studies in Health Technology and Informatics*, *264*, 413–417.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–887.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*, 230–238.

Wang, F., & Preininger, A. (2019). AI in health: State of the art, challenges, and future directions. *Yearbook of Medical Informatics*, *28*, 16–26.

Wasylewicz, A. T. M., & Scheepers-Hoeks, A. M. J. W. (2018). Clinical decision support systems. In *Fundamentals of clinical data science* (pp. 153–169). Springer International Publishing. https://doi.org/10.1007/978-3-319-99713-1-11

Weegar, R., & Dalianis, H. (2015). Creating a rule based system for text mining of Norwegian breast cancer pathology reports. In *Proceedings of the sixth international workshop in health text mining and information analysis, louhi 2015, held in conjunction with emnlp 2015*, Lisbon, Portugal.

Weegar, R., Nygård, J. F., & Dalianis, H. (2017). Efficient encoding of pathology reports using natural language processing. In *Proceedings of recent advances in natural language processing, ranlp 2017*, Varna, Bulgaria.

White, C. M., Schmidler, G. D. S., Butler, M., Wang, Z., Robinson, K., Mitchell, M. D., ... Banez, L. (2017). *Understanding health systems' use of and need for evidence to inform decisionmaking* (Tech. Rep.). https://doi.org/10.23970/ahrqepcwhitepaper2

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *25*, 1419–1428.

Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, *19*(S5), 1–9. https://doi.org/10.1186/s12911-019-0935-4

Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). *Transfer learning*. Cambridge University Press.

Ye, W., Hu, R., & Enev, M. (2020, August). Put Deep Learning to Work:Accelerate Deep Learning through Amazon SageMaker and ML Service. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. ACM. https://doi.org/10.1145/3394486.3406698

Yu, S. (2016). Big privacy: Challenges and opportunities of privacy study in the age of big data. *IEEE Access*, *4*, 2751–2763. https://doi.org/10.1109/access.2016.2577036

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zhang, Y., Cai, T., Yu, S., Cho, K., Hong, C., Sun, J., Huang, J., Ho, Y. L., Ananthakrishnan, A. N., Xia, Z., Shaw, S. Y., Gainer, V., Castro, V., Link, N., Honerlaw, J., Huang, S., Gagnon, D., Karlson, E. W., Plenge, R. M., ... Liao, K. (2019). High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nature Protocols*, *14*, 3426–3444.

How to cite this article: Tayefi M, Ngo P, Chomutare T, et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comput Stat*. 2021;e1549. https://doi.org/10.1002/wics.1549