The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges

TABINDA SARWAR, SATTAR SEIFOLLAHI, JEFFREY CHAN, XIUZHEN ZHANG, VURAL AKSAKALLI, IRENE HUDSON, KARIN VERSPOOR, and LAWRENCE CAVEDON, RMIT University

The primary objective of implementing Electronic Health Records (EHRs) is to improve the management of patients' health-related information. However, these records have also been extensively used for the secondary purpose of clinical research and to improve healthcare practice. EHRs provide a rich set of information that includes demographics, medical history, medications, laboratory test results, and diagnosis. Data mining and analytics techniques have extensively exploited EHR information to study patient cohorts for various clinical and research applications, such as phenotype extraction, precision medicine, intervention evaluation, disease prediction, detection, and progression. But the presence of diverse data types and associated characteristics poses many challenges to the use of EHR data. In this article, we provide an overview of information found in EHR systems and their characteristics that could be utilized for secondary applications. We first discuss the different types of data stored in EHRs, followed by the data transformations necessary for data analysis and mining. Later, we discuss the data quality issues and characteristics of the EHRs along with the relevant methods used to address them. Moreover, this survey also highlights the usage of various data types for different applications. Hence, this article can serve as a primer for researchers to understand the use of EHRs for data mining and analytics purposes.

CCS Concepts: • **Computing methodologies** → **Information extraction;**

Additional Key Words and Phrases: EHR, data types, data characteristic, data challenges, data mining, health analytics

ACM Reference format:
Tabinda Sarwar, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. 2022. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Comput. Surv.* 55, 2, Article 33 (January 2022), 40 pages.
https://doi.org/10.1145/3490234

Tabinda Sarwar and Sattar Seifollahi contributed equally to this research.

Sattar Seifollahi is currently working at Resolution Life (Australia). This work was performed while working at RMIT University.

This work was partially supported by funding from Telstra Health and the Digital Health Cooperative Research Centre, which is funded through the Australian Government's Department of Industry, Science, Energy and Resources.

Authors' address: T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, and L. Cavedon, RMIT University, 124 La Trobe St, Melbourne, Victoria, Australia, 3000; emails: {tabinda.sarwar, sattar.seifollahi}@rmit.edu.au.

0360-0300/2022/01-ART33 $15.00
https://doi.org/10.1145/3490234

1 INTRODUCTION

Healthcare institutions (e.g., hospitals, rehabilitation centres, insurance providers, pharmaceutical developers, and aged-care facilities) regularly record the health data of clients/patients in digital systems referred to as **Electronic Health Records** (**EHRs**). This data consist of heterogeneous elements, including demographics, prescriptions, diagnosis, vital signs, immunizations, laboratory and radiology test results, medical concepts and notes, procedures, and treatment plans. This information is recorded each time a client/patient visits a hospital or healthcare organization. Overall, EHR systems improve the quality of healthcare services by:

- enabling data sharing across multiple healthcare organizations such as research laboratories, specialists, medical imaging facilities, pharmacies, emergency facilities, and medical schools;
- allowing access to real-time data and tools to help healthcare providers in decision-making about patient's care plans;
- better data tracking over time;
- automating the workflow for clinicians;
- providing timely reminders for patient screenings and preventative checkups to improve patient care;
- facilitating research by providing medical history and related healthcare data of the patients.

While the primary goal of EHR systems is efficient and effective management of health information [1], current healthcare research and practice have become more data-driven and evidence-based in medical assessment, diagnosis, treatment, and prevention. The EHR encapsulates extensive data for a large population of patients that provide both the healthcare and research communities the capacity to perform effective retrospective research and analytics to improve the well-being of people. Thus, the availability of EHR databases [2–5] has increased the opportunities for secondary usages [6, 7].

EHRs have been extensively used in the last 10 years for various clinical and research applications. Many machine learning algorithms such as logistic regression [8], Naive Bayesian [9], support vector machines [10], random forests [11], and neural networks [12, 13] have been employed for mining the EHR data. These methods find their application in tracking the progression or trajectory of a disease, cohort identification, health risk prediction, and adverse event detection [14]. Though the successes and promising results of data mining methods have been reported in the literature, it should be noted that raw EHR records suffer from a variety of data challenges, limitations, and quality issues that must be addressed prior to developing any data-driven models.

This review is dedicated to the preliminary step of understanding the information stored in the EHR data (Figure 1), as understanding the types of data and its associated characteristics affect the quality of data-driven models and research. It should be noted that many studies have proposed and analyzed the characteristics and dimensions of data quality to measure and evaluate the *fitness for use* of EHR data [6, 15–20]. The main focus of this article is to comprehensively understand EHR data in the context of data mining. We review the factors, characteristics, limitations, and challenges that potentially affect the quality of data mining processes and the relevant methods that are used to address them. Furthermore, we conducted a comprehensive analysis on the abstracts of over 1,336 papers published between 2010–2020 to study the relationship between the data mining applications and various data types, along with the EHR challenges addressed by the current studies. Hence, this article is one of the most comprehensive surveys in understanding the use of EHRs for data mining, reviewing the domain knowledge of various data types and associated challenges, and also discusses the methods used in the literature to address these challenges.

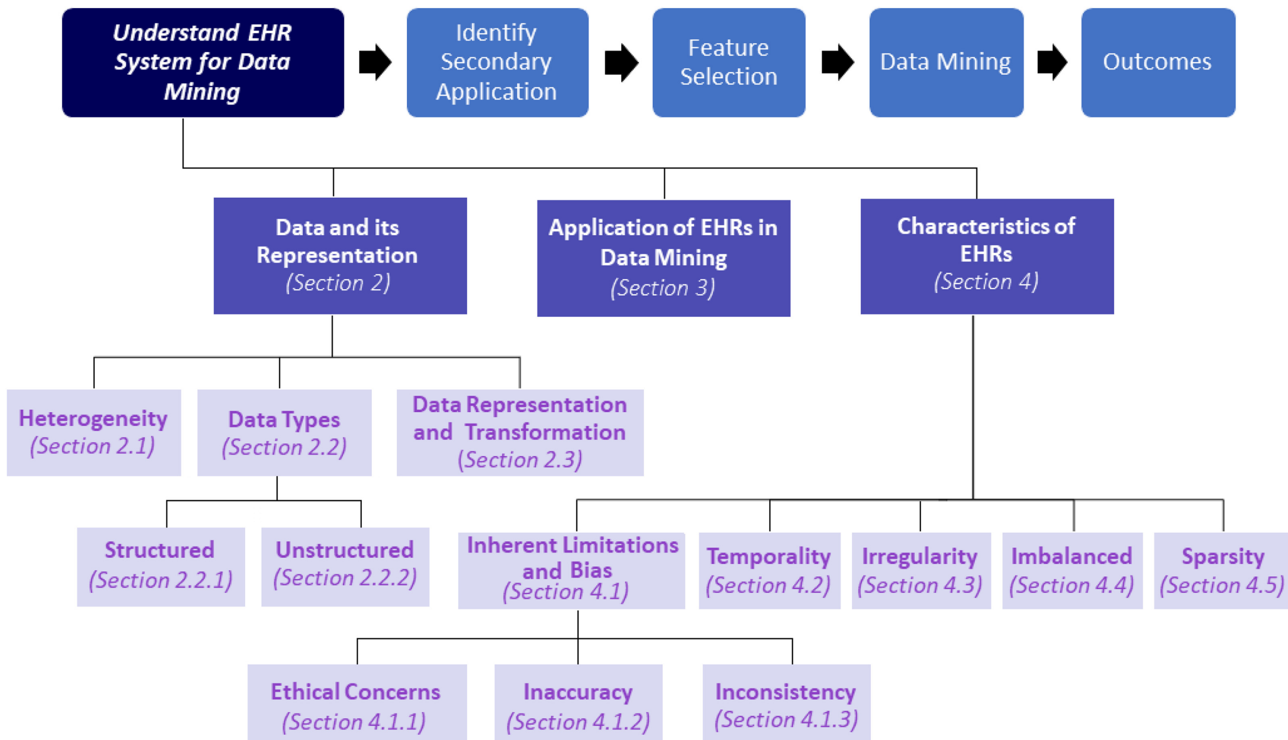

Fig. 1. The general framework of using EHRs for secondary use. The preliminary and essential step is to understand the EHR data as it is associated with various data types, applications, characteristics, and limitations. After understanding the EHRs domain, secondary application of EHRs is identified, followed by a selection of relevant features and data mining strategy. Based on the secondary application, the outcomes from the data mining can then be used either for clinical decision-making or as research findings.

It should be noted that data mining models and techniques associated with EHRs are not covered in this survey. For reviews on the application of data mining in the EHRs, the readers are referred to [14, 21–29]. Yadav et al. [14] reviewed the study design (cohort, case-control, cross-sectional, and descriptive studies) for using EHR data, along with the usage of data mining methodologies for major clinical applications (e.g., understanding the natural history of a disease, cohort identification, risk prediction). Jensen et al. [21] also discussed the collection and application of health data for various applications including genetics and genomics. The review of data mining techniques for specific applications such as phenotyping [26], adverse drug detection [27] and coronary artery disease [28] can also be found in the literature. Solares et al. [22] and Shickel et al. [25] reviewed the state-of-the-art deep learning models applied to EHR data. Esteva et al. [23] further discussed the application of different deep learning strategies for various medical and clinical applications including medical imaging, robotic-assisted surgery, and genomics. Luque et al. [24] specifically reviewed the text mining techniques for medical applications like medical concepts extraction, text summarization, text classification, and so on. Stiglic et al. [29] provided an overview of interpretable prediction models used in the healthcare domain.

In this review, we discuss various data formats commonly used in EHRs, their associated characteristics, and challenges of using EHRs in data mining. The rest of the article is structured as follows. Section 2 introduces the heterogeneous nature of the EHR, where the data are recorded in various types. Each data type is related to specific information and provides valuable insight into the patient's health conditions. We also discuss data transformation and low-dimensional representation of EHR data, which is, in particular, important for clinical notes. Section 3 discusses research studies that have successfully utilized these data types for various healthcare applications.

In Section 4, we describe characteristics of EHR that affect the quality of the data that introduced limitations and challenges for the EHR-based research studies. Various methods and techniques found in the literature that have addressed these characteristics are also discussed in detail. Section 5 summarizes this review and broadly discusses the data and research challenges associated with EHR systems, followed by the conclusion. Thus, this article serves as a primer for researchers in the field of health analytics to understand the challenges of EHR data and relevant methodologies to address its associated characteristics and limitations. Figure 1 presents a framework of using EHR for secondary research, where the factors associated with the preliminary step (*Understand EHR Data and Identify Secondary Application*) represent an overview of this survey.

2 DATA AND ITS REPRESENTATION

As previously mentioned, an EHR contains systematized collections of patient data over time that can be shared across different healthcare settings, including health professionals and researchers. There are diverse types of data in an EHR, ranging from a patient's personal information (e.g., age, gender, and ethnicity) to medical diagnoses, prescription, and procedures performed. One of the distinguishing features of this data is its *heterogeneity*, i.e., the presence of multiple data categories for the patient; this is discussed in Section 2.1, along with the sources of heterogeneity. Information stored using two data types, structured and unstructured, are discussed in Section 2.2, followed by data transformation techniques (Section 2.3) for EHR-based research.

2.1 Data Heterogeneity

EHR data contain a wide range of information types, compared to other domains [25], and patients may have quite different information depending on their health condition. This information could be in the form of clinical observations, laboratory records, hospitalizations and discharge summaries, demographics, medications, and billing information; hence, EHRs can be considered to be heterogeneous [30]. The major forms of data within the EHR for recording healthcare information are summarized in Table 1. EHR data may represent both static and temporal information. Static data are generally recorded during patient registration process and remain stable throughout the clinical encounters, e.g., demographics information. Temporal data represent data acquired over multiple visits, which is further discussed in Section 4. For more details on various forms of data, we refer readers to [31].

The representation of EHR data is critical as heterogeneous data are generally shared between different systems, as EHR systems often interact with other decision support and financial systems to maintain the records of patients for purposes like billing and practice management. Hence, this information must be represented in a format that can ensure data standardization and interoperability between multiple applications. This is not only important for data management, but it is also to facilitate the EHR-based research studies. This is further discussed in Section 2.2.

2.2 Data Types

As discussed in Section 2.1, EHR encapsulates heterogeneous data, which could be partitioned into two broad categories: structured and unstructured data.

2.2.1 Structured Data. Structured format represents data that can take a value within a specified range or from a pre-defined dictionary. Examples of such EHR data include, but are not limited to, medical codes, medications, administrative data, vital signs, and laboratory test outcomes. Structured data can be either numeric or categorical [21]. Examples of categorical type are diagnostic, medication, and procedure codes, where the numeric data include respiration, blood pressure, pulse oximetry, and laboratory test results.

Table 1. Data Categories within EHR for Recording Patient's Health Related Information

Data Category	Description	Examples
Demographics	General characteristics of patients	Age, gender, ethnicity/race, socioeconomic status
Vital Signs	Medical signs indicating the status of the body's vital functions	Body temperature, pulse rate, respiration rate, blood pressure
Medications	Drugs and medicines, either as narratives or codes (e.g., RxNorm)	Aspirin, Potassium Chloride, Acetaminophen, Tylenol, Morphine, Buprenorphine, Valacyclovir
Diagnostic codes	Codes representing diseases and related health problems (e.g., ICD)	Acute respiratory failure - J96.00, lever lesion-K76.89, systolic heart failure - I50.2
Procedures	Medical, surgical, and diagnostic procedures, either as narratives or codes (e.g., CPT)	Eyelid skin biopsy, Partial Mastectomy, MRI Thoracic Spine
Clinical notes	Free-text written by clinical professionals (e.g., doctor, nurse, physician, radiologist) regrading patient's status	Consultation notes, discharge summaries, procedures notes, progress notes, medical notes
Laboratory data	Medical examination results, either as narratives or codes (e.g., LOINC)	Red/white blood cell count, hemoglobin, glucose, glycated hemoglobin *etc.*
Hospitalization	Data related to patient's hospital admission	Length of stay, admission source, transfer record, discharge disposition, observations

Demographics are among the most commonly captured data, including characteristics such as age, gender, ethnicity/race, marital, and socioeconomic status. The quality of some demographic features, such as age and gender, is often reasonable due to the simplicity in recording these characteristics, while they also have a significant relationship with health conditions [32]. Some features, such as a patient's income, marital, or socioeconomic status [31], may not be recorded as these are often marked as optional fields in some data collection systems. Vital signs are also widely recorded and have been successfully utilized in modelling outcomes (e.g., predicting hospital readmission [33]) and disease analysis (e.g., hypertension [34] and sepsis [35–38]). It should be noted that some research treats height, weight, **body mass index** (**BMI**), and waist circumference as vital signs [31], since they are found to be significant indicators of health and well-being: e.g., the work of [39] recommended that measurements of BMI and waist circumference should be considered as vital signs in clinical practices. However, they are not as frequently recorded as temperature and blood pressure.

A substantial amount of structured data is stored in the form of codes that standardize the representation of information. The two important standards that are applied to represent medical diagnosis and procedures are: (1) the **International Classification of Diseases** (**ICD**) codes, and (2) the **Current Procedural Terminology (CPT)**. The ICD is a globally used diagnostic tool for epidemiology, health management, and clinical purposes, which standardizes the representation of diseases, disorders, injuries, and other related health conditions. ICD codes are further divided into two categories: ICD-CM (Clinical Modification), which standardizes diagnostic codes; and ICD-**PCS** (**Procedure Coding System**), which reports the medical procedures and interventions generally used for inpatient reporting, such as for hospital billing. In the EHR literature, two versions of ICDs, namely, ICD-9 and ICD-10, are commonly used [13, 40–43]. ICD-9 and ICD-10 include around 13,000 and 68,000 diagnostic codes, respectively. Another coding system, the **Systemized Nomenclature Of Medicine Clinical Terms** (**SNOMED-CT**), which represents standardized terminology for knowledge representation in multidisciplinary clinical practice [44]. The main difference between ICD and SNOMED is that the ICD is limited to disease, while SNOMED provides a complex relationship between concepts (including clinical findings, symptoms, diagnoses, procedures, organisms, pharmaceuticals, specimens, etc).

The CPT is a standard vocabulary of codes for surgical, medical, and diagnostic procedures. While ICD is commonly used for reporting diagnoses, CPT is used for medical procedures and services. There are mainly three categories in CPTs including codes for: (1) procedures or services including evaluation and management, anesthesia, surgery, radiology, pathology, and laboratory services; (2) supplemental tracking codes used for performance measures including composite measures, patient management, patient history, physical examination, diagnostic processes, therapeutic, preventive, or other interventions, follow-up or other outcomes, patient safety, structural measures, and non-measure code listing; and (3) services and procedures for data collection, assessment and in some instances, payment of services, and procedures. The medical codes can vary between organizations and countries, with partial mappings maintained by resources such as the **Unified Medical Language System** (**UMLS**) and the SNOMED-CT. Given the large array of schemata, harmonizing and analyzing data across terminologies, and between institutions is an ongoing research [25]. The **Healthcare Common Procedure Coding System** (**HCPCS**) is another set of health care procedure code systems, to standardize the healthcare claims for health insurance providers. HCPCS utilizes the CPT to provide a standardized coding scheme to insurance providers for their billing systems. Both CPT and HCPCS are published by the **American Medical Association** (**AMA**).

Medications in the form of text strings can also be standardized using the code representations. Common vocabulary systems for standardizing medications include the **National Drug Code** (**NDC**), the **National Drug File - Reference Terminology** (**NDF-RT**), RxNorm, SNOMED, the **Anatomical Therapeutic Chemical** (**ATC**) Classification System, and a number of commercialized drug code standards such as MediSpan, Multum, **Generic Product Identifier** (**GPI**), and **First Databank** (**FDB**) [31]. The NDC is a unique product identifier for drugs intended for human use. It consists of a unique 10-digit code with a 3-segment number; e.g., the NDC for a 100-count bottle of Prozac 20 mg is 0777-3105-02. The NDF-RT is produced by the US **Veterans Health Administration** (**VHA**) as an extension of the VHA NDF by organizing the drugs into a standard representation. More specifically, it is used to model drug features such as ingredients, chemical structure, dose, physiologic effects, and related diseases. RxNorm is a US-specific terminology in medicine containing all medications available on the US market. The difference between NDC and RxNorm is that if more than one manufacturer produces the same medication, each will get different NDC value, while the RxNorm creates standard names and identifiers for the combinations of ingredients, strengths, and dose forms. The work of [45] studies NDF-RT and RxNorm for classification of medications extracted from EHRs.

Laboratory tests are also an important source of information to determine patient's overall health. Coding systems used for laboratory results include the **Logical Observation Identifiers Names and Codes** (**LOINC**), SNOMED, and CPT [31]. LOINC can be used to standardize the data from laboratory test results as well as the data from vital signs. A comprehensive list of applications that have utilized these data categories is reported in Table 2 and this will be further discussed in Section 3.

2.2.2 Unstructured Data. An exceptionally large part of EHR data is in unstructured form, which represents information recorded in the form of free-text such as clinical notes and discharge summaries. Clinical notes refer to a variety of textual documents generated on behalf of a patient in many healthcare settings. A progress note is an important sub-type of clinical notes that address a patient's health status or condition during hospitalization or over the course of outpatient care [125]. Clinical notes refer to a variety of textual documents generated on behalf of a patient in many healthcare settings. Unstructured data may include handwritten notes by healthcare providers such as admission notes, discharge summaries, medical history, procedures notes, and even notes to support management tasks like transitions of care, care planning, quality reporting, billing, outpatient visits, emergency department visits, home-care, and nursing visits.

Table 2. Research Studies between 2010–2020 That Have Utilized Different Data Categories and Features in Healthcare

Health Domain	Demographics	Vital Signs	Medications	Medical Codes	Clinical Notes	Lab Values
Dementia	[10, 46–48]	[46]	[10, 46, 47, 49]	[10, 46, 48, 49]	[10, 49, 50]	–
Falls	[51, 52]	[53]	[51–53]	[53, 54]	[54]	[52]
Mortality	[55–57]	[36, 57–59]	[55–58, 60, 61]	[57]	[57, 62]	[55, 57–59]
Sepsis Study	[63–65]	[35–38, 63–66]	[64, 66]	[67]	[64]	[37, 38, 63, 66]
Precision Medicine	–	–	[68, 69]	[69]	[69, 70]	[69]
Comorbidity	[71–73]	–	[71]	[72–75]	–	–
Phenotyping	[76–79]	[76, 79–82]	[68, 76, 77, 79–84]	[41, 75, 77–79, 81, 82, 84–91]	[30, 41, 79, 80, 84, 88–90, 92]	[76, 79, 82, 88]
Suicide	[93–98]	–	[94–97]	[93, 99–101]	[99, 101]	–
Depression	[98, 102, 103]	–	[102–104]	[105]	[104, 105]	–
Readmission	[8, 33, 57, 73, 106, 107]	[33, 57, 108]	[8, 33, 57, 107–109]	[33, 57, 73, 106, 107, 110]	[33, 57, 107, 111–113]	[8, 33, 57]
Kidney Injury	[79, 114–116]	[79, 108, 114]	[79, 114]	[79, 114, 116, 117]	[79]	[79, 116, 118]
Diabetes	[47, 119–121]	[119]	[44, 47, 61, 119, 121–123]	[88, 120, 122, 124]	[88, 121, 121, 125]	[44, 88, 119–124]
Heart Study	[78, 126, 127]	–	[60, 126, 128, 129]	[78, 88, 126, 127, 129]	[88, 127, 130, 131]	[88]
Cancer	[56, 132, 133]	–	[56]	[132–134]	[133–136]	[133]

Extracting knowledge from notes is challenging [137] due to one or more of the following challenging issues: (1) non-standardized formats; (2) abundant typing and spelling errors; (3) violation of natural language grammar; and (4) existence of abbreviations, acronyms, and idiosyncrasies.

2.2.3 Miscellaneous Data. As well as structured or unstructured data formats, another category of data is semi-structured, which has not always been well-understood by the research community. This data type does not reside in fixed fields/records i.e., many healthcare organizations use semi-structured data to record custom and non-standardized information. Some of EHR data related to this category might be flowsheets or drop-down menus in EHR systems. An example of such data is "name" with the corresponding "value" for a laboratory test result such as "blood pressure". For more detail, we refer the reader to [14]. We also refer the reader to [138], where some semi-structured data such as **JavaScript Object Notation** (**JSON**) and **eXtensible Markup Language** (**XML**) are discussed in the EHR context. It can be argued that such data are more similar to the structured data as its value is usually restricted, unlike clinical notes.

It should be noted that other medical modalities such as ultrasound, radiographs, **Magnetic Resonance Imaging** (**MRI**), **electrocardiogram** (**ECG**), and so on are also types of unstructured data, which are recorded for patients in hospitals associated with their health condition. Different medical modalities have been used in isolation to structured and textual data for diagnostics in dermatology, radiology, ophthalmology, and pathology [23]. Each modality requires specialized pre-processing methods and background knowledge for data mining and comprehension of results, so studying the characteristics of these modalities is beyond the scope of this review. For -Omic data such as genomics, transcriptomics, epigenomics, proteomics, metabolomics, and integration to EHR data, we refer readers to [137].

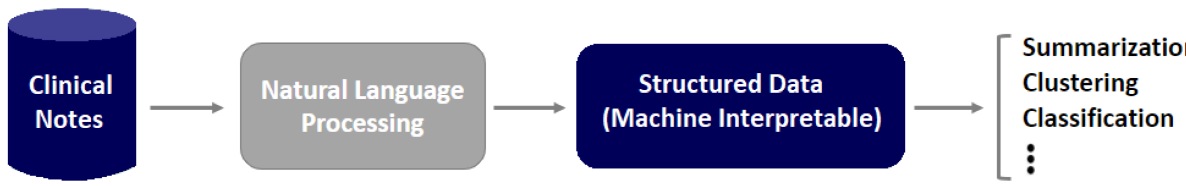

Fig. 2. Data processing pipeline for clinical notes.

2.3 Data Representation and Transformation

Data preparation and transformation are important stages in the data analytics pipeline, where substantial data processing is performed before applying machine learning and data mining algorithms. The performance of these algorithms heavily depends on the quality and type of the data. Data, which are not primarily collected for research purposes (such as EHR data sourced from hospitals), need cleansing and transformation before it can be used for data analysis. One of the main tasks in data preparation is to transform the EHR data into a study design matrix so that data mining techniques can be effectively applied to develop solutions for specific clinical applications. For a detailed explanation of different types of study design for matrix transformation of the EHRs, we refer readers to the work of Yadav et al., [14].

Transforming unstructured data into a reliable representation space (referred to as embedding) is also of substantial importance as data mining algorithms cannot be directly applied to raw text. The following section reviews some methodologies used to transform the EHR data for mining purposes. We will also discuss the low-dimensional representation of the data, which is important for clinical notes and categorical high-dimensional structured data, e.g., vocabulary size in the text corpus of clinical notes or the total number of ICD codes.

2.3.1 Transformation of Clinical Notes. Textual data, such as progress notes and discharge summaries, contains rich and important information that cannot be easily captured using structured data. However, text data need to be converted into structured (numeric) features for data mining purposes. Figure 2 shows a pipeline of transforming raw textual information into data suited to the performance of analytics, where NLP techniques play a key role in data processing. Research advancement in NLP provides high-performance solutions to data-driven models. **Bag-of-Words** (**BoW**) is a traditional and widely-used representation for texts [112, 125]. This model represents a text instance as a vector of word with corresponding frequencies. One alternative approach to BOWs is to use UMLS **Concept Unique Identifiers** (**CUIs**) in the medical domain for better representation and annotation of medical textual data [90]. The **term frequency-inverse document frequency** (**tf-idf**) is another traditional technique to convert a text corpus to a mathematical but high-dimensional representation. Several NLP systems have been developed to facilitate the task of standardising textual information in clinical settings. Some of the frequently used systems are reported in Table 3.

Several advanced neural network-based embedding techniques like Word2Vec [153], **Embedding from Language Models** (**ELMo**) [154], **Robustly Optimized BERT approach** (**RoBERTa**) [155], and **Bidirectional Encoder Representations from Transformers** (**BERT**) [156] can successfully extract context from the text (semantic and syntactic similarity), unlike standard BoW models. Some research studies that have utilized these neural networks to transform the free-text data in EHRs include: ClinicalBERT for transforming clinical notes predicting hospital readmission [112]; a topical word embedding in deep learning called EnHANs to annotate patient's notes with ICD-9 codes [157]; extracting diagnosis from free-text data by using semi-supervised machine learning [158]; a multi-layer representation learning, Med2Vec, for representing medical concepts [13], and cui2vec for embedding medical concepts [159]. It should also be noted

Table 3. NLP Systems for Standardizing Textual Information

NLP System	Description
cTAKES	Open-source system using the UIMA framework that extracts clinical data with contextual attributes like polarity and certainty and generates structured output using SNOMED-CT, UMLS, and RxNorm [45, 139–147]
MedKATp	A pathology extraction system that uses rules to map text to elements of the Cancer Disease Knowledge Representation Model [143, 148, 149].
MedLEE	Rule-based system for structuring radiography reports and expanded for nearly all types of clinical notes and waslater commercialized [99, 142–145, 150]
MetaMap	A originally designed for literature abstracts that assigns the best candidate UMLS terms to segments of text and can map output to any constituent terminology in the UMLS [143, 145, 148, 151]
OpenNLP	An Apache project for NLP that includes components like a sentence boundary detector, tokenizer, symbol remover, and POS tagger, as well as Max Entand Perceptron named entity recognizers [141, 146, 152]
SymText/MPLUS	A system with Bayesian Network-based semantic grammar that can extract and normalize findings from radiography reports [143–145]

that neural networks could be used for reducing the dimensionality of data, which is discussed in Section 2.3.2.

2.3.2 Low-Dimensional Representation. As previously discussed in Section 2.1, the high-dimensionality of the EHR data is due to its heterogeneous nature. Moreover, the data are sparse as patterns of patient health conditions and healthcare vary between patients and among visits. For unstructured data, sparsity can also stem from the fact that each patient has a limited vocabulary or set of diagnostic codes associated with the clinical notes, while the vocabulary dictionary is comparatively large. For example, using the BoW model for such scenarios will result in a high-dimensional vector, which might lead to poor performance of data mining algorithms. Hence, a low-dimensional representation of data is valuable for facilitating data analysis.

Dimensionality of the data to be analyzed is a challenge in EHR, as the number of features can be extremely large, e.g., UMLS contains more than 210 biomedical vocabularies with over 2.4 million concepts [160]. A common approach to address high-dimentionality is to remove highly correlated and frequent features from the feature space, particularly for the case of the BoW model. However, this may not result in substantial dimension reduction. To further reduce feature space dimension, there are two common strategies: feature selection and feature reduction. Feature selection is the more popular approach for structured data, whereby features can be manually selected for a specific health application or can be selected using machine-learning algorithms (e.g., random forests). Examples of feature selection include feature selection methods based on confidence and information gain [161], knowledge sources based on the rank correlation between the concept of the target phenotype and other candidates [89], and usage of **correlation feature selection** (**CFS**) to identify a subset of features highly correlated with the outcome and weakly correlated amongst themselves [162].

A large category of dimensionality reduction techniques is related to feature reduction and building a new lower-dimensional feature space. For example, Garg et al. [163] transformed the progress notes into 150 features for ischemic stroke subtype classification. There are two main categories of feature reduction include: (1) traditional techniques such as **Principle Component Analysis** (**PCA**) and **singular value decomposition** (**SVD**); and (2) deep learning techniques e.g., (Med2Vec [13], BERT, etc).

Latent Semantics Analysis/Indexing (**LSA/LSI**) and **Latent Dirichlet Allocation** (**LDA**) are NLP-based techniques that have been used for feature reduction for textual data [49, 164]. LSA adopts SVD to identify patterns or relationships between the terms and concepts in a text corpus

[42]. LDA, on the other hand, is a topical model applying a probabilistic mixture model to cluster related words together, thereby reducing the dimensionality of the text representation. Although embedding techniques have been primarily developed for text data, there is growing interest in using these techniques on structured data (e.g., medical concepts) for low-dimension representation of a such data [42, 82, 106]. Some examples of research studies applying deep learning in EHRs include: Word2Vec for low-dimensional representations of medical concepts from the structured EHR [82]; word embedding on medical codes guided by prediction task [42]; embedding medical codes into a unified vector space for EHR phenotyping [41]; embedding with raw text and CUI for phenotyping [79]; embedding medical entities into a harmonized space, by utilizing both structured and unstructured sources [43]; "deep patient" to automatically represent patients based on a set of general features, through a deep learning approach [165]; and an attention-based bidirectional **recurrent neural network** (**RNN**), called diagnosis prediction model (Dipole), for learning low-dimensional representations of medical concepts [166].

3 APPLICATIONS OF EHRS IN DATA MINING

Research studies have utilized various categories of EHR data (reported in Section 2.2) to solve important health and clinical problems. Structured data have been used for many health applications, e.g., phenotyping [78, 87, 91, 167], diabetes detection [61, 123, 124], mortality prediction [55, 59], and cancer diagnosis [132]. Similarly, clinical notes have also being used for similar applications including geriatric syndrome [168], dementia [10, 49, 168], mortality prediction [57, 62, 69], heart problem studies [127, 130], diabetes [125], and hospital readmission [112]. The mentioned studies used only either structured (specific data categories) or unstructured, but a combination of both can also be found in the literature (Table 5) e.g., Shao et al. [49] used a combination of structured and unstructured data to diagnose dementia through a weakly supervised machine-learning approach.

3.1 Literature Analysis

We conducted a comprehensive abstract analysis of research papers on EHR. The aim of this analysis is to understand the trends in EHR research, i.e., which data types and data mining strategies are frequently used for EHR data. We also want to analyze the usage of various data types for different healthcare applications. For this purpose, we utilized abstracts of the 1,336 papers, which were published between 2010–2020 with the focus of data mining for EHR only. The papers were selected using a combination of search terms that included "electronic health record" or "ehr" with "data mining", "machine learning", and "deep learning" in PubMed data.[1]

We used cosine similarity to quantify the relationship between data types and health applications. To calculate this relationship, we first defined a list of important n-grams or keyword phrases (here 110 n-grams). An n-gram is a sequence of n words from a given text content, which serve as potential features in text analytics [127]. Some of the n-grams are used in Figures 3 and 4, where $n = 1$ (unigrams) or $n = 2$ (bigrams). We selected a list of n-grams related to our focus categories, also considering their frequencies in research papers. More precisely, they are selected from three different categories: (1) n-grams related to data, e.g., "clinical notes", "medications", and "laboratory values"; (2) n-grams related to health applications; e.g., "mortality", "hospital readmission", and "precision medicine"; and (3) n-grams related to methodology, e.g., "neural networks", "support vector machines" and "natural language processing". For each n-gram, we found a list of corresponding papers (abstracts) in which the n-gram occurs frequently, and concatenated all corresponding papers (abstracts) to build a single text document. The title and author-provided

[1] https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

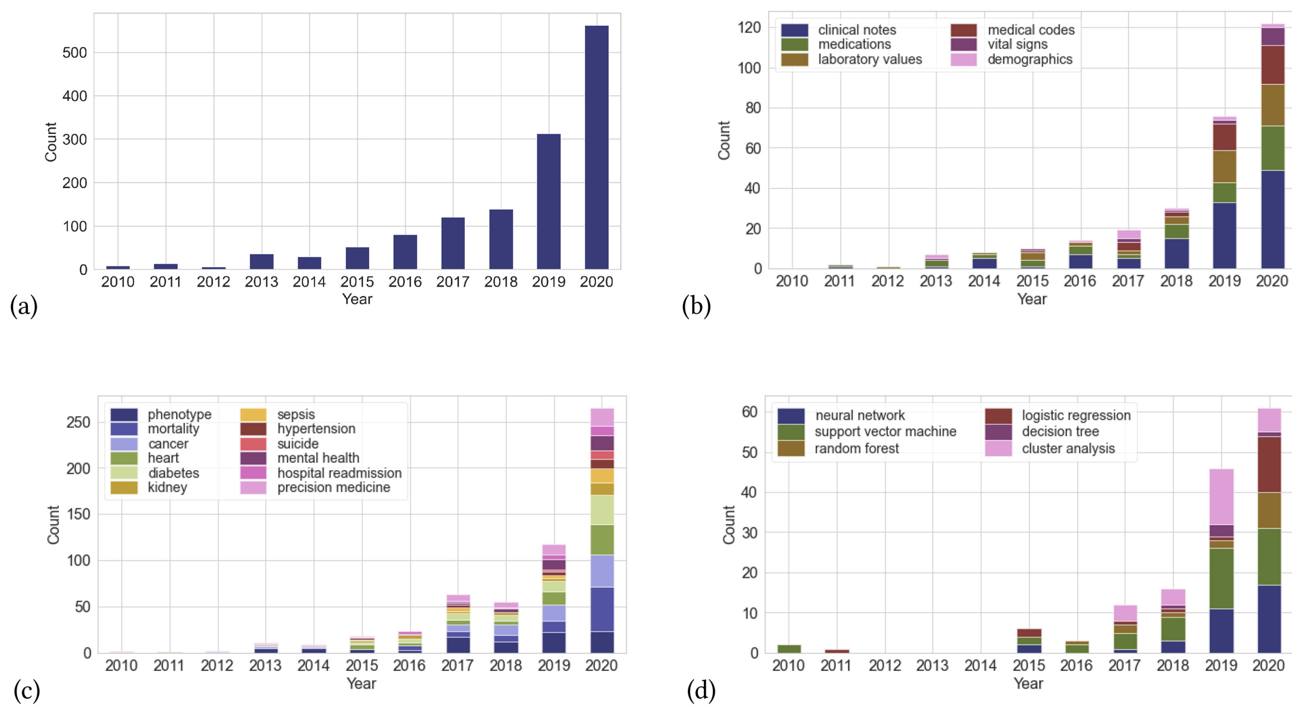

Fig. 3. During 2010–2020, (a) the number of publications (b) types of data used (c) health applications, and (d) methodology used for secondary applications of EHRs. X-axes show the year, while y-axes demonstrate frequency of papers. It is noted that overlaps between various items in each sub-figures of (b)–(c) altered the scaling along the y-axes.

keywords of papers were also included in the analysis. Later, we computed the most frequently occurring tokens and bi-grams of each text document. We then calculated tf-idf on text documents to convert them to numeric vectors. Cosine similarity was then used to find relationships (similarity) between pairs of n-grams. This analysis is presented in Figure 4(a), where it shows association of various data types to different health applications, in particular, the importance of unstructured data (clinical notes) to health applications.

3.2 Trends in Research

Table 2 reports the detail of the data categories used to address the various healthcare applications. It is evident from Table 2 that phenotyping is one of the widely studied research topics utilizing EHR data. Vital signs have been frequently used for mortality, sepsis, hospital readmission prediction, and heart study. Medications are widely linked with health problems such as dementia, falls, mortality, phenotyping, suicide, depression, hospital readmission, kidney, diabetes, and heart disease. Clinical notes have been commonly utilized in recent years for many applications involving dementia, mortality, precision medicine, phenotyping, hospital readmission, diabetes, heart studies, and cancer. The most common medical research applications in the context of EHR includes phenotyping, hospital readmission, mortality prediction, kidney disease, diabetes, and heart failure. For more in-depth review of applications of data mining utilizing EHR data, the reader is referred to other studies in the literature [14, 16, 21–26, 162, 169].

In Figure 3, the number of publications, data used, health applications, and methodology used are analyzed over the years 2010–2020, which highlights the popularity of topics in recent years. The number of publications nearly doubled in 2020 compared to 2019, and the same for 2019 compared to 2018 (Figure 3(a)). Our analysis unveils stable growth in the number of publications prior to 2018. This highlights the recognition of data mining in resolving healthcare research problems. Figure 3(b) shows data used in research papers (abstracts) in last 10 years. An increase in the use of

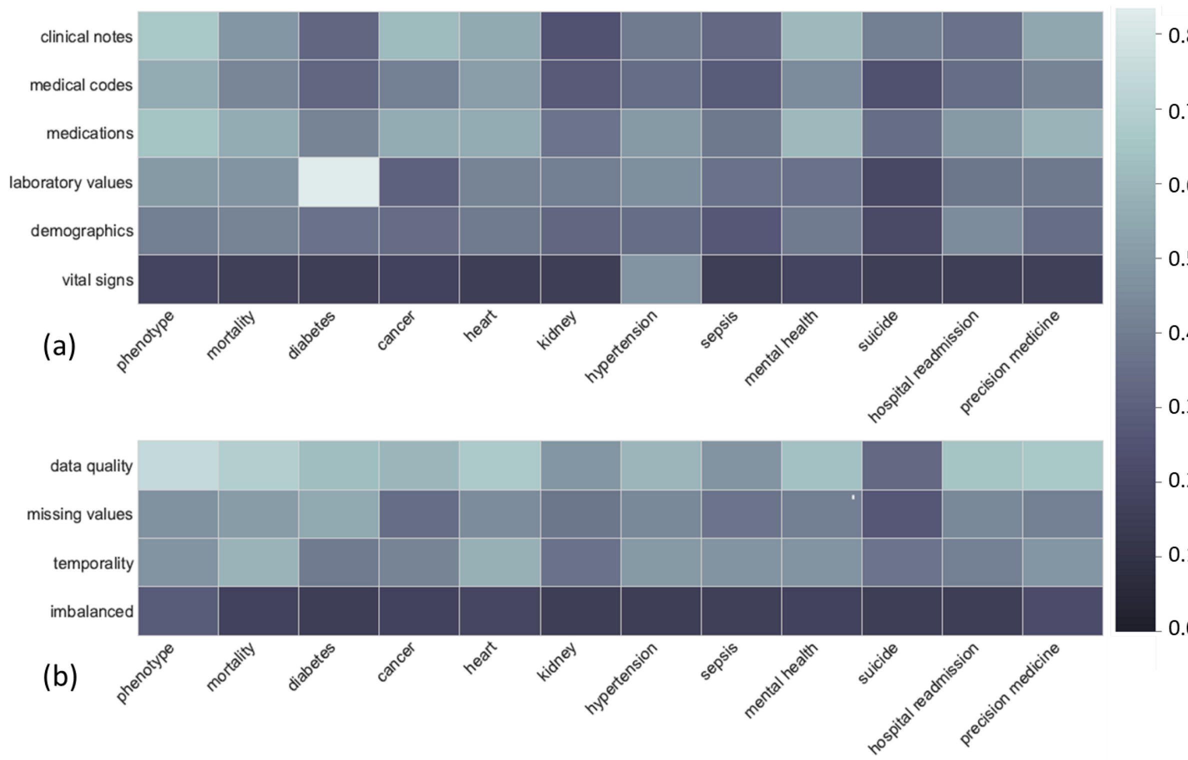

Fig. 4. Analysis of research literature on EHR studies published between 2010–2020. This included computing cosine-similarity between various healthcare applications and (a) data categories, (b) EHR characteristics.

clinical notes stems from recent development in the state-of-the-art technologies such as BERT and the proven high predictive ability of textual data. For health applications, the trend seems similar across different health applications with a significant increase in analysis applications in recent years due to the increase in the number of publications; see Figure 3(c). Figure 3(d) represents the use of different data mining methods in the last 10 years, which shows that deep learning has gained popularity in recent years for health analytics. It should be noted that some of the research papers have not explained the data used, the methodology, or the health application in their abstract that resulted in different scaling of y-axes.

From Figure 4(a), one can also observe relationships between data types and health applications. Some of the most frequent features used by various studies, reported in Figure 4(a), include laboratory values, medications, demographics, medical codes (ICD, CPT, SNOMED, etc) and vital signs. Among these data types, medications, medical codes, and clinical notes have been well utilized in phenotyping, mortality, cancer, precision medicine, and mental health. Some applications are not very well-studied in the context of EHRs, e.g., kidney, diabetes, sepsis, and suicide. But the interest in EHRs has grown in recent years due to major developments in data mining techniques. Hence, future studies could focus on addressing heterogeneous health problems. Figure 4(b) focuses on the relationship between characteristics of EHRs and health applications, which are discussed in Section 4.

4 CHARACTERISTICS OF EHRS

Patients visit a hospital when they require medical assistance. The required assistance vary for different patients, depending on their physical and medical needs. Moreover, the time duration between consecutive visits also varies. Such factors result in numerous characteristics of EHR data that makes the data mining process quite challenging. These characteristics include temporality, irregularity, data imbalance, and sparsity, which are discussed in Sections 4.2–4.5. Figure 5

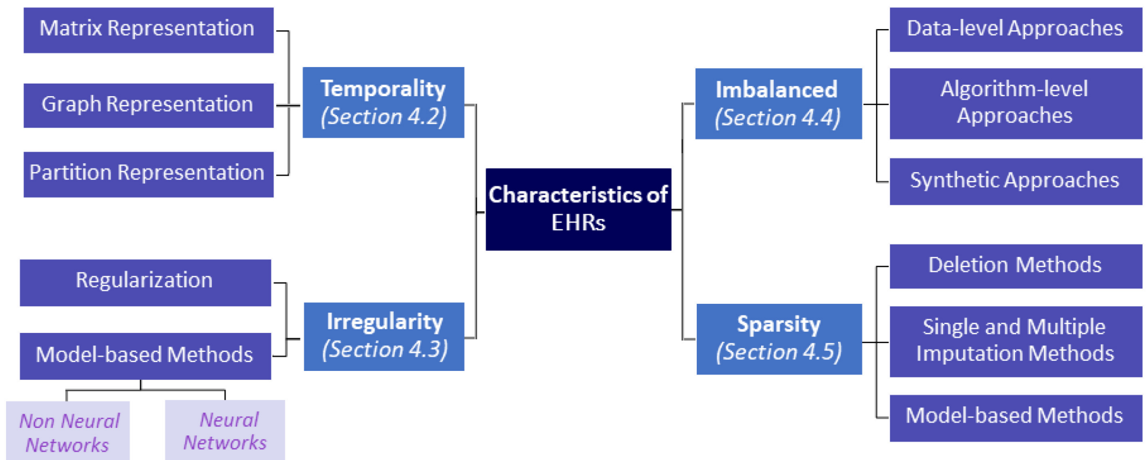

Fig. 5. An overview of the characteristics of EHRs and relevant strategies for addressing them.

summarizes strategies for addressing the characteristics of the EHR, where these strategies have been successfully utilized by many research studies (Table 5).

Before we discuss these characteristics, Section 4.1 discusses some of the bias and errors associated with the EHR, which are important considerations when building an EHR data-driven model. We also analyze the relationship between the characteristics of EHR data and health applications (Section 3.2), where we find that a limited number of studies have addressed these characteristics (Figure 4(b)).

4.1 Inherent Issues in EHRs

Data for observational studies are typically curated and generally considered to be of good quality, i.e., accurate, and free of bias [15, 170]. But EHR data are not primarily collected for research studies. Moreover, the protocols for recording research data and real-time health data also varies, so it is possible that the data quality needs for a research study are not satisfied. A major concern in the secondary usage of EHRs is that as the dataare not systematically collected for research, errors can be introduced anywhere in the process from observation to conceptualization of the patient's results [171]. This can have negative impact on findings extracted from the dataset [6] and loss of predictive power [137]. Here, we discuss limitations encountered during data curation, thus referred to as inherent limitations, along with the ethical concerns associated with EHR (Section 4.1.1). These issues are summarized in Table 4.

4.1.1 Ethical Concerns. Primary and secondary uses of EHRs provide benefit for wider society (patients, physicians, clinicians, and researchers) and provide an opportunity to discover unrecognized risk factors (e.g., health deterioration or mortality prediction), but also raises ethical concerns. The important ethical issue of using EHR data for research include privacy and data security [172], informed consent for data uses [173], and ownership of patient data [174]. To address such issues, EHR policies and systems employ numerous techniques that include legal requirements, encryption, data de-identification, access limits, and audit logs in order to protect data privacy [174]. Generally, breach of data privacy can result in both civil and criminal liabilities [175]. Moreover, the data mining of EHRs should also ensure public benefit, i.e., research that contributes toward improving the public health system. Ensuring fairness in the data mining model is also critical as sensitive features such as gender, age, race, and sexual orientation should not bias the decision-making process in the healthcare domain [176].

A detailed review on the ethical practices is beyond the scope of this review, but we refer interested readers to [177] on a survey on ethical and regulatory frameworks for the provision of

Table 4. Summary of Inherent Limitations of EHRs

Issues	Explanation
Ethical Concerns	Includes issues of data privacy, security, informed consent for data uses, ownership of patient data, ensuring public benefit, and unbiased decision-making processes (fairness in the data model).
Data Inaccuracy	**Erroneous Data:** EHR suffers from data entry errors e.g., selection of incorrect menu items in the medical system, replication of clinical notes from prior visits, the time difference between the actual time of data collection and electronically recording the same data, *etc*.
	Software Constraints: The quality of the data recorded depends on the EHR software packages, e.g., pseudo-examination (predefined patient-relevant questions) offer convenience to the clinician for recording medical history but may not accurately reflect the patient's signs and symptoms.
	Loss of Information: Patient information could be lost due to issues like data fragmentation (data recorded at multiple hospitals) and poor recording of disease classification.
	Data Biases: EHRs is prone to many biases which include selection bias, confounding bias, information bias, survival bias, etc.
Inconsistency	Inconsistent data representations in the datasets (e.g., presence of multiple data formats, units or measurement protocols for recording same data) or inconsistencies in various records (e.g., difference ICD codes for the same patient).

mental healthcare, and [176, 178, 179] on ethical frameworks and challenges for developing artificial intelligence technologies within clinical-based contexts and research.

4.1.2 Data Inaccuracy. Data mining of EHRs relies on the assumption that the recorded data are accurate. But this assumption is not always applicable to real EHR datasets as they typically contain data that are erroneous, incomplete, miscoded, and fragmented due to factors such as clinician workload, time limitation, or poor user interfaces of the EHR systems [169]. A few of the data-related problems encountered in EHRs include

— **Erroneous Data:** EHRs often suffers from data entry error mistakes because sometimes clinicians type quickly, click incorrect menu items, or may replicate the clinical notes from prior visits without carefully reviewing the content [169]. It is also common that the actual time of data collection is different from the time at which it is electronically recorded. It often happens that clinical notes are not recorded by a single person, e.g., for in-patients, multiple medical staff such as nurses, physicians, clinicians, and so on are involved in the data collection and recording process. Generally, a high-level clinician has the responsibility to confirm the findings in the notes; however, to save the time, some may not read the notes carefully to confirm accuracy of the data [180]. Bias is also observed for other data modalities like waveform data, where the common quality issues include random noise, gaps in the waveform, and artifacts (e.g., patient's motion) [137].

— **Software Constraints:** Van Der Bij et al. [181] observed a substantial difference in the quality of recording among clinical reports and software packages. The study reported that 30%–100% of healthcare episodes had a meaningful diagnostic code, which depended on the EHR software package used for recording data. The study recommended standardizing the functionalities of the EHR software packages to improve the quality of data. Many EHR systems provide the functionality of creation and usage of note templates to save the time of clinicians. The pre-defined template could be modified to incorporate the details of individual patients. The risk involved in templated notes is that time-poor clinicians may carelessly extract fragments of the normal examination findings from the template that were not observed or assessed during the medical examination [180]. This phenomenon was reported by Bernat [180], where the templated endoscopic reports from one particular physician were identical for several patients. Also, many EHR systems provide pre-defined patient-relevant questions with an option of either yes or no. Although these binary questions, referred to as pseudohistory or pseudo-examination, offer convenience to the clinician for recording medical history, patient's symptoms have degrees of variability, subjectivity, and changeability [180]. Hence, these standardized questions may not accurately reflect the patient's signs and symptoms.

- **Loss of Information:** Clinical records are often fragmented, e.g., a patient might consult multiple clinicians in different hospitals. Generally, the EHR systems across multiple hospitals do not communicate with each other and often the systems are not interoperable [169]. This fragmentation causes information loss, which can lead to inaccurate research outcomes. While data are recorded at a healthcare institution, the information could be generalized leading to the loss in the granularity of the details. Botsis et al. [182] reported information inaccuracy in EHR, where the granularity of the diagnosis or disease classification code was not reflected in the records.
- **Data Biases:** For any research study, a well-defined standard is developed for the selection of eligible subjects for the study. For EHR-based study, the criteria could be age limitation, missing clinical information, a limited number of clinical visits, or presence/absence of a medical condition. Hence, a strict selection criterion affects the validity of the analysis and its applicability to a broad population (referred to as selection bias) [183].

  Confounding bias is also generally observed in analyzing EHRs, where the health status of different patients can affect the true relationship or lead to spurious outcomes [184]. Consider an example where one patient has diabetes but now is suffering from depression. Confounding factors may falsely demonstrate a false association between the two medical conditions. Generally, confounding bias is observed when the distribution of a known prognostic factor differs between two groups [185]. Other biases that can affect the outcomes of a EHR-based study include information bias [186], admixture bias [187], incidence-prevalence bias [188], survival bias [189], and treatment bias [190].

4.1.3 Inconsistency. The consistency could be violated due to the presence of multiple data formats, units, measurement protocols, and granularity [191]. Data granularity refers to the degree of details required to record a feature. Many individuals are involved in the data collection process, such as nurses, physicians, clinician, and so on, which could lead to inconsistent data representations in the dataset. The secondary usage of EHRs may require manual processing to assess data quality and standardize data format for analysis [18, 192].

Information inconsistency in EHRs was reported for documentation where chemotherapy regimes were recorded in clinical notes instead of the drug register for patients with pancreatic cancer [182]. Inconsistencies were also observed between various records, e.g., pancreatitis was diagnosed as chronic in the pathology reports but recorded as acute in the clinical notes. Moreover, inconsistencies were observed within the record of the same patient, where two different ICD codes were recorded for the same patient.

Studies in the literature rely on the assumption that the EHR dataset was accurate and consistent, which is why these studies have not addressed these properties as a limitation. To our knowledge, the framework proposed by Lee et al. [43] is the only study that addressed inconsistency in diagnostic code assignment in the EHR dataset. The study proposed a unified graph representation learning framework to embed heterogeneous medical entities (structured and unstructured datasets) into a harmonized space. The method demonstrated high accuracy in detecting erroneous diagnosis codes, which were introduced artificially to the data for evaluation purposes.

4.2 Temporality

EHR data are inherently temporal. Patients may seek care repeatedly based on the individual's health condition and needs. The number of visits and duration between the consecutive visits

Table 5. Characteristics of the EHR Dataset That Have Been Successfully Utilized and Addressed by Data Mining Studies

Papers	Unstructured	Structured	Temporality	Sparsity	Irregularity	Imbalanced
[8, 11, 96, 102, 162, 193–195]		✓				
[119]		✓				✓
[34, 76, 116]		✓		✓		
[196]		✓		✓		✓
[77, 166, 197–202]		✓	✓			
[40, 85, 107, 120]		✓	✓			✓
[58, 203, 204]		✓	✓		✓	
[63, 205, 206]		✓	✓	✓		
[12, 13, 41, 80, 98, 207–216]		✓	✓	✓	✓	
[217]		✓	✓	✓		✓
[57]	✓	✓	✓	✓		
[9, 79, 121]	✓	✓		✓		
[10, 49, 168]	✓	✓				
[69, 218]	✓	✓	✓	✓	✓	
[219]	✓		✓	✓	✓	
[62]	✓		✓			✓
[220–223]	✓					

vary for every patient. Observations such as temperature or blood pressure could change over time for a patient, making time-series analysis appropriate. The combination of variation in the observations make it difficult to directly compare different patients and requires pre-processing for a fair comparison. Furthermore, patients who require more medical attention may have more frequent visits than less sick patients. This implies that on the visit-level, the data are biased as multiple samples from the same patient will demonstrate inter-dependency, which can affect the analysis outcomes. One possible way to address this issue is to use models that can deal with non-independent data, e.g., mixed effect models [224]. The time irregularity between the visits and features are discussed in Section 4.3.

Sequential medical events over time provide valuable information on the trajectory of the health condition that could be used for disease detection, prediction, and progression. Thus, EHRs collected over time provides opportunities for longitudinal analyzes for diverse research problems. To address temporality, the dataset features require transformation to account for multiple visits recorded in the EHR data. For this purpose, three types of data representation have been used in the literature.

— **Vector representation:** The most common method to capture temporality is to partition the records into vectors for each visit, i.e., a feature set for each visit is generated separately [12]. Some studies compute mean, median, standard deviation, minimum, and maximum value of the vectors from multiple visits [76, 162, 225], which results in loss of temporal information. This loss could be addressed by using an RNN or similar technique, where the medical features of a visit are merged with the previous visit's features. Studies that have used vector representations for training RNNs include [13, 41, 107, 166, 203, 208–210, 212, 213, 215–218, 226]. Many studies have also introduced attention networks to interpret the importance of temporal features. For example, Liu et al. [58] introduced two event attention mechanisms to identify critical events and temporal dependency of different events.

— **Tensor representation:** The dataset is transformed into a tensor where three dimensions represent the visits, selected features, and number of patients. Afshar et al. [77] represented temporal features in the form of a matrix $x^k \epsilon \mathbf{R}^{I^K \times J}$ where I^K is the number of clinical visits for patient and K and J are the total number of medical features. Hence, a tensor representation comprised these temporal matrices for all the patients in a cohort. Similar tensor representation of temporal features was also utilized in [12, 205, 207, 214].

— **Graph representation:** A sequence of events could also be represented in the form of a graph [80, 227]. In graph representations, a node consists of the medical heterogeneous features and a directed edge represents the sequence of the visits. The graph is generally weighted, i.e., a weight is computed for the edges with the rationale that a smaller weight is computed for larger time intervals between the event nodes, as it is less likely that the temporally distant events are related. Thus, weighted graphs could be used to address both temporality and irregularity (Section 4.3). Hettige et al. [40] used a bivariate graph for feature set representation. Specifically, the graph had two nodes partitions, visits and diagnosis codes, where edges denoted a link between the visit and diagnosis code.

4.3 Irregularity

Unlike the time-series data where observations are recorded at regular intervals, EHR is longitudinal data as the recorded observations for the patients are irregularly recorded i.e., it suffers from irregular time intervals between patient's visits. Generally, there are two levels of irregularities observed in EHR data, namely, visit-level and feature-level [216]. Visit-level irregularity refers to the irregularity observed in the patient's visit, as patients visit the hospital when clinical care is required. Thus, the frequency of the visit and the interval between the visits can not be standardized. Feature-level irregularity refers to the appearance of the same feature irregularly in the EHR dataset. For example, vital signs are collected at every visit, while laboratory tests are conducted after a certain time-gap or when required [14]. Generally, feature-level irregularity is addressed as sparsity, which is further discussed in Section 4.5. The rest of the section is dedicated to studies that have addressed visit-level irregularity during data mining.

Time stamps for the health care record are critical for assessing the health condition of the patient. For example, a patient may not visit a hospital for months or years, but later develops a disease that requires frequent clinical visits. Moreover, irregularity also provides valuable insights into disease progression [210, 216, 230]. The credibility of analysis and the performance of the data mining algorithms could be affected if the irregularity in the dataset is ignored. The same phenomenon was reported by Hripcsak et al. [228], who evaluated four types of time parametrizations for irregular EHR. The study reported a decrease in predictability of glucose measurements with the increasing time gap between the measurements (Figure 6(a)).

Many data mining studies assume that data are sampled at regular time intervals, where the time-gap between the visits was ignored. The majority of the studies discussed in Section 4.2 relied on this assumption. These studies considered each visit to be equally important, although it may happen that visits with a wider time-gap are not as important as others. However, it is not plausible to consider healthcare records as regular longitudinal data. Consideration of regular intervals in EHR reduces the distinction between features with long- and short-term dependencies. In order to

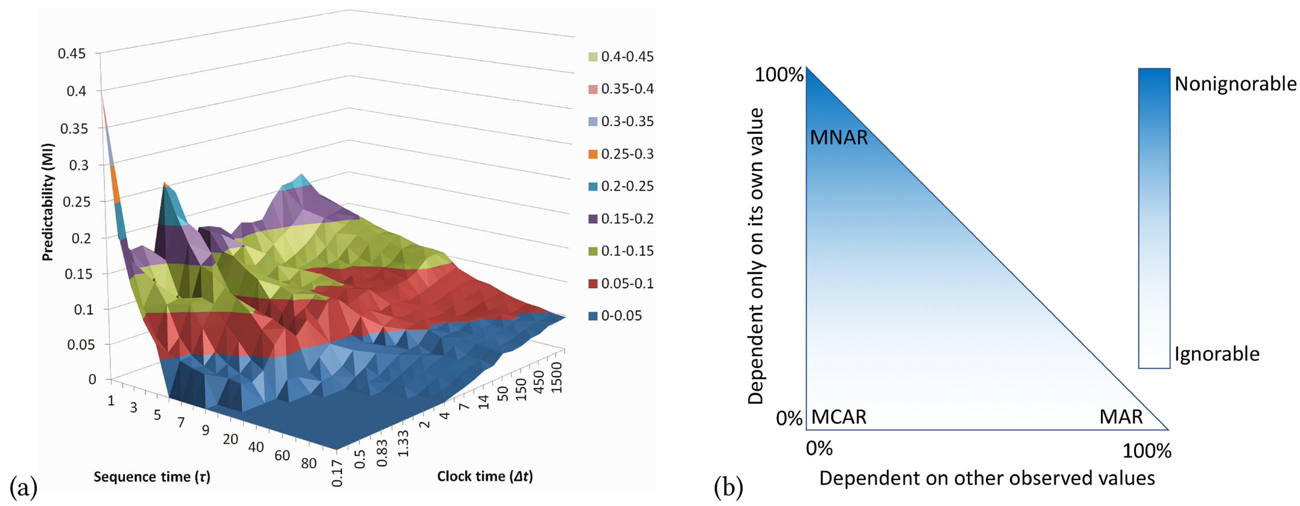

Fig. 6. (a) Predictability of glucose plotted against sequence time (number of measurements) and clock time (time interval between two measurements). Predictability is highest at the shortest clock and sequence time and it drops with the increasing clock time. Also, there is a dramatic drop in predictability with increasing sequence time. Reproduced from [228] (b) the interchangeability of MCAR, MAR, and MNAR assumptions. The x-axis indicates the extent to which a given value being observed depends on other values of other observed variables. The y-axis indicates the extent to which a given value being observed depends on its own value. Reproduced from [229].

address irregularity, the time stamps need to be included in the data analysis. Two solutions so far exist in the literature to address irregular timestamps: (1) Convert irregular data to regular time intervals (regularization), and (2) modify the model to incorporate irregular timestamps.

— **Regularization:** Liu et al. [58] used adaptive segmentation to address temporal irregularity for clinical outcome prediction, where the records were segmented based on the time difference between successive visits, such that the multiple segments have visits with regular time intervals. Alternatively, Gupta et al. [211] aggregated the visits within a specified window to regularize the dataset.
— **Model-based methods:** For accurate modeling of an EHR dataset, many studies have modified and developed models to address the irregular timestamps.
  – *Non-Neural Networks:* The Smith–Waterman algorithm [231] was modified in one study to compute the temporal similarity between irregular laboratory tests, where the time difference between two observations was directly incorporated in the similarity metrics [232]. Time warping was also introduced for phenotype discovery using irregular samples of record [193]. Escudié et al. [75] used Sperrin's coefficient [233] to regularize the records for studying autoimmune comorbidities in patients with celiac disease. In another study, the time stamps for the visits were integrated with structured support vector machines to detect and monitor the progression of disease [234]. The **Drug Effects on Laboratory Test** (**DELT**) method modelled the time variation to detect drugs that have effects on laboratory tests [214]. Marginal and conditional models can also handle irregular data by assuming a correlation among multiple observations of a patient with irregular time gap [14].
  — *Neural Networks:* RNNs are well-known for capturing the time-dependence for sequential data, where this data could be a sequence of words or events. To address irregularity, RNNs have been successfully used to handle the sequence of irregular patient's visits. Baytas et al. [208] proposed a **long short-term memory** (**LSTM**) network to extract patient

subtypes from the EHR dataset. The study proposed a novel time-aware unit to learn the time decay for addressing irregular time intervals encountered in the longitudinal patient record. Time-lapse was also integrated in the memory cell of LSTM to account for irregular visits [215, 218]. Specifically, the time decay and time parameterization was introduced on the forgot gate of the LSTM unit. Che et al.[203] utilized the concept of time warping to measure similarity between two longitudinal records to model the gate parameters in the proposed **Gated Recurrent Unit** (**GRU**)-based 2D RNN for predictions of Parkinson's disease.

A time-aware convolutional network, a combination of RNN and **convolution neural network** (**CNN**) layers, integrated the time stamps into the convolutional layer [226]. The weights of the layer were adjusted according to the time stamps with the assumption that temporally close events are more relevant for disease prediction. Zhang et al. [227] developed a heterogeneous CNN for comorbidity risk prediction in which the records were transformed into graphs, where the diagnoses were used as nodes and the edges were computed from the temporal intervals. A similar methodology was used to generate graph for temporal phenotyping [80]. The time interval between the consecutive visits was embedded in the longitudinal input vector for many neural networks-based studies [13, 40, 41, 113, 212, 216, 219, 230, 235].

4.4 Imbalanced Data

EHR data is dominated by class imbalance, where the class of interest is heavily underrepresented in comparison to the other classes [236]. A large volume of patient information with a wide range of disease are routinely recorded [237] and as a consequence, the dataset may be dominated by non-disease patients (normal cases) or less-severe cases. This can result in poor discrimination and calibration of data mining models [237], as only a small percentage of patients suffer from severe or chronic disease and most patients have symptoms of milder manifestation of disease [236].

Many of the input features may be labeled with different probabilities to different classes, referred to as *overlapping classes*, [238], i.e., similar sign and symptoms can be observed for multiple diseases. The overlap between a rare and prevalent class and an imbalanced dataset can cause the trained model to associate the data points with the prevalent class, resulting in poor sensitivity toward the rare class [238]. Thus, not addressing the imbalanced class distribution in the dataset can result in a model with biased and poor performance. This poor performance has many associated health risks, e.g., a biased classification may increase the risk of health deterioration or overuse of medication in patients. Hence, the inhomogeneous distribution of classes needs to be addressed for the successful and error-free deployment of disease prediction and detection models.

As class distributions are highly skewed in EHR datasets, accurate detection and representation of the rare class are important, as these classes may correspond to high impact events. For many cohort-specific studies [8, 9, 11, 12, 40, 106] the data are carefully selected, which eliminates the risk of biases introduced by imbalanced classes. But for other applications, there exist three broad strategies for handling imbalanced data [239]. It should be noted that many studies did not explicitly address the imbalance of data, but used evaluation metrics such as **area under receiver operating curve** (**AUROC**) and **area under precision recall curve** (**AUPRC**) to demonstrate accurate performance of the methods under imbalanced dataset [13, 22, 40, 77, 85, 217, 237, 240]. These studies are not reported in this section and Table 5.

— **Data-level approaches:** This involves randomly *oversampling* the rare class, *undersampling* the prevalent class, or a combination of both strategies. The oversampling strategy involves replication of the data instances, i.e., multiple copies of the same data points exist in the dataset. Alternatively, the undersampling strategy involves removing data points of the prevalent class to match its number to the number of rare class instances. These are the most popular techniques for addressing the imbalanced data due to its simplicity and computational efficiency [211, 229, 239, 241–245]. But there are drawbacks associated with each of these strategies. Oversampling can lead to the problem of *overfitting* as it involves replication of data points, resulting in the creation of specific rules [246]. A loss of information is involved in undersampling as some data points are ignored in the analysis.

— **Algorithm-level approaches:** In data-level approaches, the dataset is pre-processed to address the class imbalance before it can be used for analysis. But several studies have incorporated a rebalancing mechanism in the algorithm to deal with inhomogeneous class distribution. This is a common strategy for deep learning methods, where the loss function was modified to address the imbalanced dataset. Zhu and Razavian [196] introduced a penalty for false predictions generated by imbalanced data, i.e., the loss function is weighted by inverse class weight for each outcome node of the neural network. Qiu et al. [120] introduced the cost of misclassification as a cost ratio of the rare class against the prevalence class to rebalance the class distribution. Graph Laplacian priors were applied for training a neural network to classify the physiologic time series with imbalanced diagnostic labels [85].
— **Synthetic approaches:** This category consists of algorithms that specifically address the imbalanced data by generating data points for the rare class. **Synthetic Minority Oversampling Technique** (**SMOTE**)[247] generates the samples using a linear combination of two samples x and x^r from the rare class, where x^r is sampled from the K-nearest neighbors of sample x. SMOTE has been successfully used in various studies [248–251]. The **adaptive synthetic** (**ADASYN**-[252]) model extends SMOTE by using a weighted distribution for different minority classes; i.e., more data samples are generated for minority classes that are harder to learn as compared to easier minority classes. Similarly, a synthetic data generation algorithm has been proposed for both rare and prevalent classes [253]. Jian et al. [254] instead generated a synthetic dataset using a pairwise combination of data points and class labels, and their exclusive disjunction was used to generate modified labels for the dataset.

4.5 Sparsity

EHR data are also characterized by its sparse nature, i.e., it consists of many missing values. There are many reasons for sparsity. Due to variability in individual medical needs, it is not necessary that the same information is recorded for each patient. For example, a patient suffering from mental illness will have different assessment from a diabetic patient. Moreover, information recorded at each visit for a patient may not be the same. If a patient was scanned via MRI in one visit, it is not necessary that the same scan will be recorded on later visits as well, or it may happen that the patient might be recommended an **electroencephalogram** (**EEG**) scan instead. These two factors of variability in the information recorded for the patients introduce sparsity in the EHR, affecting its data quality. Other cases of missingness (missing data values) relates to the data collection process, which also includes lack of documentation, i.e., the observed data were not recorded in the EHR system. Missing values can also arise from the lack of data integration between hospitals. A patient can consult multiple doctors at different hospitals but usually the EHR systems do not communicate information, which could also lead to missing data. This condition is often ignored in the research. Censoring is also a common type of missingness in time-to-event analysis (e.g., survival and event history analysis). This could be a consequence when individuals withdraw before the end of study (right-censoring), or the event of interest occurs before the start of the individual is included in the study (left-censoring), thereby the data points do not exist for such cases. Interval-censoring can occur when the event of interest happened within a certain period of time but the exact information in unknown.

Missing data presents various problems, including complications in the data analysis, potentially leading to biased outcomes of the analysis, reduction in the statistical power (i.e., the probability that the test will reject the null hypothesis when it is false), and biased representation of the data samples [255]. It should be noted that the extent of sparsity or data quality depends on the intended research problem, i.e., the quality of the dataset depends on the features/variables of interest identified for a specific problem [15]. Sparsity is generally addressed as missing values that require consideration of many factors [256]:

— How many records or variables have missing values?
— Does a relationship exist between the characteristics of the feature and its value?
— Does the missingness of one variable affect the missingness of other variables?

To deal with sparsity for structured data types, the missing value problems are generally divided into three types [256–258]:

— **Missing Completely at Random (MCAR):** The records included in the analysis are not different to the excluded records, i.e., the probability of missingness is the same for all the records.
— **Missing at Random (MAR):** The missing records may depend on the observed records. Under the MAR assumption when the subgroups are created for known data values, the missingness of a variable is not systematically different from the known value within a subgroup. Thus, the probability of being missing is the same only within groups defined by the observed data.
— **Missing Not at Random (MNAR):** When MCAR and MAR assumption both do not apply, then it is characterized as an MNAR problem, i.e., missingness is related to observed and unobserved records.

The interchangeability of the three assumptions are summarized in Figure 6(b). The simple MCAR assumption is often considered unrealistic and leads to biased estimates [257]. In the case of the MNAR assumption, the only way to obtain an unbiased estimate of the variable or feature is to model the missing data [255]. In the case of EHRs, the relationship between different variables is generally expected; hence, it seems reasonable to assume the MAR assumption [259]. It should be noted that if the *Missingness Assumption* (e.g., MAR) does not hold for the dataset, it could result in a biased analysis [260].

The methods used to deal with sparsity are categorized into three types [261]:

— **Deletion Methods:** The traditional approach for addressing the missing data involves the deletion of incomplete records. This could be further divided into two subcategories. *Listwise deletion*, also referred to as complete case analysis, involves removing the cases that do not have all the relevant features/variables [262]. Complete case analysis holds validity if the data satisfies the MCAR assumption [259]. If MCAR does not always hold for a dataset, this could result in biased estimates. Even if MCAR applies to the dataset, the deletion leads to loss of information. On the other hand, *pairwise deletion* (also known as available case analysis) attempts to reduce the data loss observed in complete case analysis [263]. This involves the deletion of cases relating to each pair of variables with missing data. This usually involves computing correlation or co-variances among variables to identify deletion cases. Deletion methods have been successfully utilized by many studies for data preparation [8, 75, 96, 106, 196, 211].

— **Single and Multiple Imputation Methods:** Imputation methods estimate the missing values to avoid the loss of information observed in deletion methods. Single imputation methods are those that estimate the single value from the dataset [264], which could be in the form of the mean or median value of the observed data values. Sometimes in the case of longitudinal data, the missing value is replaced by the last or next available observation, referred to as **Last Observed Carried Forward** (**LOCF**) or **Next Observation Carried Backward** (**NOCB**), respectively. Other single imputation methods include linear interpolation, hot deck, and cold deck methods [261].

  Multiple imputation method involves creating $m > 1$ datasets for the observed data to estimate the missing value. Individual datasets are used to estimate the missing value either using a single imputation or predictive methods. The final result involves pooling from the m estimated values [261]. The studies that have used imputation methods include [85, 197, 200, 203, 208, 210, 213, 216, 217, 225, 236, 237, 265].

— **Model-based methods:** This category involves a predictive model such as regression, maximum likelihood, random forest [266], or neural network [9, 11, 80, 204, 205, 209, 213, 225, 236, 240, 265] to estimate the missing values. To address the missingness due to data censoring, approaches like non-parametric, semi-parametric, or parametric models can be used for time-to-event analysis [14].

There does not exist a standard imputation method for EHR and the selection of the method depends on the research problem and choice of the researcher. Generally, it can be assumed that the studies that have analyzed specific cohorts have used deletion methods for addressing sparsity, as it involves a selection of a subset from the complete dataset [12, 13, 249, 267–270]. Here, we have discussed only those studies that have explicitly reported the imputation method. It should be noted that sometimes missing data can be informative and could be incorporated with the data mining model [271] e.g., including an additional parameter for indicating the missing values. But this informative missingness depends on the transportability of the missing data mechanism, which can be compromised if the missing values become known. Readers are referred to [255, 256, 264, 272–275] for a comprehensive review on imputation methods for EHR data.

5 DISCUSSION

EHRs store data of individuals who visit healthcare institutions, e.g., hospitals, rehabilitation, insurance providers, pharmaceuticals, and aged-cared facilities, where the primary purpose is to efficiently manage information and data related to patient's conditions. EHRs have been widely adopted for various secondary uses, which include but are not limited to cohort analysis, phenotyping, disease classification, progression, and prediction. Data for specific research goals and observational studies are usually curated with great care to satisfy the needs of the research problem. As the primary goal of the EHRs is the efficient management of data and medical history of patients, its secondary usage poses many limitations and challenges. In this survey, we have reviewed data types, data transformation, inherent limitations, and characteristics of the EHR data that might pose multiple challenges for the researchers.

EHRs consist of a wide range of information which, in general, include demographics, medical history, prescriptions, diagnosis, vital signs, immunizations, laboratory test results, medical notes, procedures, and treatment plans. This vast collection of information is generally recorded in two major types of data, namely, structured and unstructured. A majority of EHR features, e.g., vital signs and laboratory tests, are recorded using structured data format. On the other hand, a large amount of information such as disease signs and symptoms, patient's allergies, precautionary measures, and so on are encapsulated in clinical notes. Due to this diverse nature of the dataset, EHR data are referred to as a heterogeneous and high-dimensional (Section 2.1), which poses various challenges and limitations for researchers (Section 4). It is also evident that some types of data have been more frequently used in research than others such as medications and demographics, due to their ease of use, and availability (refer to Figure 4). In recent years, the usage of free-text notes has gained much attention in clinical research. Processing clinical notes is challenging due to their subjective nature and lack of standard protocols for recording this data. But as clinical notes is a common mode to capture the critical information that could not be captured by structured data (e.g., physiological condition of patients), recent research trends have focused on developing specialised NLP algorithms for extracting valuable information from them (Table 5). Using a combination of both structured and unstructured data can also be observed in the recent research studies(Table 5), but a standard pipeline or guidance on using this data could not be found in the literature.

Generally, many secondary applications of EHR data assume that the data quality requirements are met. But unfortunately, violations of the data quality dimensions (e.g., completeness, correctness, granularity) are observed in EHR data [17]. Hence, it is important to evaluate the quality of EHR data before using it for secondary applications. For this purpose, data quality models that not only assess the structural conformance and completeness of data but also the semantic (mapping clinical concepts to data variables) quality could be used to verify its *fitness for use* prior to any analysis [18].

A well-defined standard or guideline for data pre-processing, transformation, and preparation could not be found in the literature because such techniques typically depend on the specific research application and study design [276]; e.g., quantifying the effect size of treatment will have a different data pre-processing pipeline than mortality prediction. Moreover, none of the current studies have evaluated the effectiveness of pre-processing pipelines, e.g., evaluating the performance of different embedding and transformation techniques of the progress notes for the same clinical application. Hence, there is still no “best” way of pre-processing or transforming the data.

As previously mentioned, the inter-individual variability in the medical and healthcare needs of the patients introduces many challenges for the EHR-based research (Section 4). Although some studies reported in this survey have successfully addressed one or more potential characteristics of the EHR data, the challenging nature of EHRs is evident from Table 5, as not all the characteristics were well-addressed by current studies. We also computed the association between EHR data types, characteristics, and healthcare applications (refer to Section 3 for details), which is presented in Figure 4(b). It is evident that a few of these characteristics were addressed by specific applications. The traditional method of handling the characteristic challenges involves the combination of multiple techniques (e.g., imputing missing values, regularizing the data with irregular time gaps) to pre-process the EHRs, which could later be used for mining purposes. But the focus of recent studies involves addressing these challenges by developing sophisticated models (Table 5).

Generally, there does not exist any agreed framework to address the characteristics of EHRs. The traditional solution for imbalanced data (e.g., SMOTE technique) and sparse data (e.g., mean value, LOCF and NOCB) targets structured data for classic feature-based supervised problems [261]. It should be noted that the sparsity of the structured data is very well studied in the literature, while this remains an open challenge for free-text data. The recent advanced neural network studies have addressed these challenges (Table 5) and have also demonstrated resistance to imbalanced data without directly addressing this data challenge (measured in terms of AUROC and AUPRC [22, 217, 237]). However, these methodologies pertain to specialized research problems and cannot be generalized for other studies. So there does not exist any standardized framework to address the EHR data challenges and the current solutions apply to specific research problems. Due to these limitations and variability in the study designs, the effectiveness of methods to address the characteristics of the techniques could not be independently evaluated. Salgado et al. [261] studied the performance of various data imputation methods and recommended that multiple imputation approaches have comparatively better performance than single and model-based (regression and K-nearest neighbor) imputation approaches, but it should be noted that a limited number of imputation methods were analyzed in the study. Studies evaluating the performance of techniques addressing temporality, irregularity, and imbalanced data could not be found in the literature. So, it can be concluded that the selection of the data processing method depends on the research application, study design, available computational capabilities, and preferences of the researchers.

EHR systems hold a huge amount of patient data with diverse health conditions. The distribution of diagnosed disease is generally skewed, i.e., a fewer number of mild medical conditions are reported frequently (e.g., cold, flu, and fever) as compared to severe health conditions, where these mild medical conditions are typically not a focus of research studies. Generally, the research problems and outcomes are limited to a few specific diseases such as heart disease, diabetes, cancer, and so on. (refer to Table 2 and Figure 4). The majority of the current studies have not demonstrated the accuracy of their methods on a broad range of diseases [277], due to the constraints on data availability. Though studies have demonstrated the potential of EHRs in improving our healthcare system, there is a need to target a broad range of diseases for the general benefit of individual's health utilizing limited available data.

For the data mining of EHRs, structured (e.g., diagnostic codes, vital signs and medications), and textual (e.g., clinical notes) data have generally been utilized, ignoring the inclusion of other data modalities [278] such as ECG, MRI, radiographs, and so on. These modalities have been used individually in isolation to structured and unstructured data for diagnostics in dermatology, radiology, ophthalmology, and pathology [23]. These modalities encapsulate important biological information that could be used to identify the disease biomarkers in data mining [279]. The combination of structured/unstructured datasets with other medical modalities could potentially provide superior results with biological or clinical significance. It is worth noting that the majority of the current EHR mining methods assume that the data models can capture the human physiology and pathophysiology without integrating any domain knowledge into the models [14]. This might introduce the risk of capturing a factor that may conflict with the domain knowledge [280].

For secondary application of the EHRs, the accuracy, interpretability, and trustworthiness of the data models are a major concern for real-world applications in medicine [281–283]. Generally, deep learning methods have demonstrated promising potential in research but interpretability of the models has always been a debatable topic. But identifying the underlying features or symptoms of health deterioration is a major concern for medical research. In the past few years, this need has recognized the importance of interpretable neural networks and studies have attempted to decrypt the results of neural network models. RNNs offer interpretability and trustworthiness, reducing the barriers of implementing the machine learning model in the clinical practice [14, 237].

In the context of understanding the EHRs and associated characteristics, future work should focus on developing methods to identify and rectify the inaccuracy and inconsistency found in the EHRs, which is often ignored by current studies. Moreover, developing guides and standards for data preparation methods under different study designs are also essential [14]. This can be achieved by performing systematic comparisons of the current state-of-the-art pre-processing and transformation methods under different clinical applications. This will also aid new researchers to adopt well-established methods with known performances. Moreover, the comparison of performance and predictive power of structured and unstructured data should also be studied to identify the applications where one data type can outperform the other [284]. Currently, the performance of methods for addressing the various characteristics is also unknown. Studies comparing data mining models for the same dataset can be found in the literature [22], but it is important to identify the optimal methods to address the characteristics of EHRs for different research applications and study designs.

6 CONCLUSION

Secondary usage of EHRs for research has increased dramatically in the last decade, resulting in discoveries that improve the well-being of individuals and support the decision-making process for medical stakeholders, e.g., physicians, clinicians, nurses. In this article, we have reviewed data types, biases, and characteristics of EHR data that can serve as a primer for data mining researchers who intend to utilize EHR data for their studies. EHRs contain rich information of patients in the form of various data types, including both structured and unstructured. These types have been used either individually or in combination for health applications such as disease prediction, detection, progression, cohort analysis, and phenotyping. As EHR data are not recorded solely for research purposes, there are, in general, errors and inconsistencies in the data. Moreover, characteristics of the EHR data, such as temporality, irregularity, sparsity, and data imbalance, introduce various challenges for data-driven research. Although these characteristics have often been successfully addressed in the literature to solve complex medical research problems, there does not exist any standard framework that can be used as a guide for EHR-based research. The current methods for addressing these characteristics can depend on the specific data and mining model selected for the research problem, which reduces the generalizability of these methods. This review provides a comprehensive discussion on the current methods for addressing data types and challenges in health applications of EHRs, which can be used as guidelines for future data mining studies.

REFERENCES

[1] Lawrence L. Weed. 1968. Medical records that guide and teach (concluded). *Yearbook of Medical Informatics* 212 (1968), 1.

[2] Emily Herrett, Arlene M. Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd Van Staa, and Liam Smeeth. 2015. Data resource profile: Clinical practice research datalink (CPRD). *International Journal of Epidemiology* 44, 3 (2015), 827–836.

[3] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (2016), 1–9.

[4] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 2015. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* 12, 3 (2015), e1001779.

[5] Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data* 5 (2018), 180178.

[6] Michael G. Kahn, Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G. Johnson, Siaw-Teng Liaw, Marianne Hamilton-Lopez, Daniella Meeker, C. Toan Ong, Patrick Ryan, Ning Shang, Nicole G. Weiskopf, Chunhua Weng, Meredith N. Zozus, and Lisa Schilling. 2016. A harmonized data quality assessment terminology and framework for the secondary use of Electronic Health Record data. *eGeMS* 4, 1 (2016).

[7] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and S. S. Iyengar. 2016. Computational health informatics in the big data age: A survey. *Comput. Surveys* 49, 1 (2016), 1–36.

[8] Efrat Shadmi, Natalie Flaks-Manov, Moshe Hoshen, Orit Goldman, Haim Bitterman, and Ran D. Balicer. 2015. Predicting 30-day readmissions with preadmission Electronic Health Record data. *Medical Care* 53, 3 (2015), 283–289.

[9] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. 2017. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access* 5 (2017), 8869–8879.

[10] Zina Ben Miled, Kyle Haas, Christopher M. Black, Rezaul Karim Khandker, Vasu Chandrasekaran, Richard Lipton, and Malaz A. Boustani. 2020. Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine* 102, 2020 (2020). DOI: http://dx.doi.org/10.1016/j.artmed.2019.101771

[11] Deepak Agrawal, Cheng-Bang Chen, Ronald W. Dravenstott, Christopher T. B. Strömblad, John Andrew Schmid, Jonathan D. Darer, Priyantha Devapriya, and Soundar Kumara. 2016. Predicting patients at risk for 3-day postdischarge readmissions, ED visits, and deaths. *Medical Care* 54, 11 (2016), 1017–1023.

[12] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 432–440.

[13] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the Advances in Neural Information Processing Systems*. 3504–3512.

[14] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining electronic health records (EHRs) a survey. *Comput. Surveys* 50, 6 (2018), 1–40.

[15] Nicole G. Weiskopf, George Hripcsak, Sushmita Swaminathan, and Chunhua Weng. 2013. Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics* 46, 5 (2013), 830–836.

[16] Kitty S. Chan, Jinnet B. Fowles, and Jonathan P. Weiner. 2010. Electronic health records and the reliability and validity of quality measures: A review of the literature. *Medical Care Research and Review* 67, 5 (2010), 503–527.

[17] Marcel von Lucadou, Thomas Ganslandt, Hans-Ulrich Prokosch, and Dennis Toddenroth. 2019. Feasibility analysis of conducting observational studies with the electronic health record. *BMC Medical Informatics and Decision Making* 19, 1 (2019), 1–14.

[18] Hanieh Razzaghi, Jane Greenberg, and L. Charles Bailey. 2021. *Developing a Systematic Approach to Assessing Data Quality in Secondary Use of Clinical Data based on Intended Use.* Technical Report. Wiley Online Library.

[19] Steven G. Johnson, Stuart Speedie, Gyorgy Simon, Vipin Kumar, and Bonnie L. Westra. 2015. A data quality ontology for the secondary use of EHR data. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2015. American Medical Informatics Association, 1937.

[20] Steven G. Johnson, Stuart Speedie, Gyorgy Simon, Vipin Kumar, and Bonnie L. Westra. 2016. Application of an ontology for characterizing data quality for a secondary use of EHR data. *Applied Clinical Informatics* 7, 01 (2016), 69–88.

[21] Peter B. Jensen, Lars J. Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 6, 13 (2012), 395–405.

[22] Jose Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H. Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, Nathalie Conrad, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Deep learning for Electronic Health Records: A comparative review of multiple deep neural architectures. *Journal of Biomedical Informatics* 101 (2020), 103337.

[23] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature Medicine* 25, 1 (2019), 24–29.

[24] Carmen Luque, José M. Luna, Maria Luque, and Sebastian Ventura. 2019. An advanced review on text mining in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 3 (2019), e1302.

[25] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A survey of recent advances in deep learning techniques for Electronic Health Record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2018), 1589–1604. DOI: http://dx.doi.org/10.1109/JBHI.2017.2767063

[26] Zexian Zeng, Yu Deng, Xiaoyu Li, Tristan Naumann, and Yuan Luo. 2018. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16, 1 (2018), 139–153.

[27] Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. *Computing Surveys* 47, 4 (2015), 1–39.

[28] Roohallah Alizadehsani, M. Roshanzamir, Moloud Abdar, Adham Beykikhoshk, Abbas Khosravi, M. Panahiazar, Afsaneh Koohestani, F. Khozeimeh, Saeid Nahavandi, and N. Sarrafzadegan. 2019. A database for using machine learning and data mining techniques for coronary artery disease diagnosis. *Scientific Data* 6, 1 (2019), 1–13.

[29] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 5 (2020), e1379.

[30] Joshua C. Denny. 2012. Chapter 13: Mining electronic health records in the genomics era. *PLoS Computational Biology* 8, 12 (2012), e1002823. DOI: http://dx.doi.org/10.1371/journal.pcbi.1002823

[31] Lehmann H., Taylor C., Ehrenstein V., Kharrazi H. Obtaining data from electronic health records. *In:Proceedings of the Gliklich RE, Leavy MB, Dreyer NA, editors. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2019 Oct. Chapter 4.* Available from https://www.ncbi.nlm.nih.gov/books/NBK551878/. ([n. d.]).

[32] Duanping Liao, Lawton Cooper, Jianwen Cai, James Toole, Nick Bryan, Gregory Burke, Eyal Shahar, Javier Nieto, Thomas Mosley, and Gerardo Heiss. 1997. The prevalence and severity of white matter lesions, their relationship with age, ethnicity, gender, and cardiovascular disease risk factors: The ARIC Study. *Neuroepidemiology* 16, 3 (1997), 149–162.

[33] Juan C. Rojas, Kyle A. Carey, Dana P. Edelson, Laura R. Venable, Michael D. Howell, and Matthew M. Churpek. 2018. Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society* 15, 7 (2018), 846–853. DOI: http://dx.doi.org/10.1513/AnnalsATS.201710-787OC

[34] Chengyin Ye, Tianyun Fu, Shiying Hao, Yan Zhang, Oliver Wang, Bo Jin, Minjie Xia, Modi Liu, Xin Zhou, Qian Wu, Yanting Guo, Chunqing Zhu, Yu Ming Li, Devore S. Culver, Shaun T. Alfreds, Frank Stearns, Karl G. Sylvester, Eric Widen, Doff McElhinney, and Xuefeng Ling. 2018. Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. *Journal of Medical Internet Research* 20, 1 (2018), e22. DOI: http://dx.doi.org/10.2196/jmir.9268

[35] R. Arnold, J. Isserman, S. Smola, and E. Jackson. 2014. Comprehensive assessment of the true sepsis burden using electronic health record screening augmented by natural language processing. *Critical Care* 18, 1 (2014), 1–10. DOI: http://dx.doi.org/10.1186/cc13434

[36] R. L. Fogerty, C. Sankey, K. Kenyon, S. Sussman, S. Sigurdsson, and A. S. Kliger. 2016. Pilot of a low-resource, EHR-based protocol for sepsis monitoring, alert, and intervention. *Journal of General Internal Medicine* (2016).

[37] Eren Gultepe, Jeffrey P. Green, Hien Nguyen, Jason Adams, Timothy Albertson, and Ilias Tagkopoulos. 2014. From vital signs to clinical outcomes for patients with sepsis: A machine learning basis for a clinical decision support system. *Journal of the American Medical Informatics Association* 21, 2 (2014), 315–325. DOI: http://dx.doi.org/10.1136/amiajnl-2013-001815

[38] Zina M. Ibrahim, Honghan Wu, Ahmed Hamoud, Lukas Stappen, Richard J. B. Dobson, and Andrea Agarossi. 2020. On classifying sepsis heterogeneity in the ICU: Insight using machine learning. *Journal of the American Medical Informatics Association* 27, 3 (2020), 437–443. DOI: http://dx.doi.org/10.1093/jamia/ocz211

[39] Robert Ross, Ian J. Neeland, Shizuya Yamashita, Iris Shai, Jaap Seidell, Paolo Magni, Raul D. Santos, Benoit Arsenault, Ada Cuevas, Frank B. Hu, Bruce A. Griffin, Alberto Zambon, Philip Barter, Jean-Charles Fruchart, Robert H. Eckel, Yuji Matsuzawa, and Jean-Pierre Després. 2020. Waist circumference as a vital sign in clinical practice: A consensus statement from the IAS and ICCR working group on visceral obesity. *Nature Reviews. Endocrinology* 16, 3 (2020), 177–189.

[40] Bhagya Hettige, Yuan-Fang Li, Weiqing Wang, Suong Le, and Wray L. Buntine. 2020. MedGraph: Structural and temporal representation learning of electronic medical records. In *Proceedings of the 24th European Conference on Artificial Intelligence* (2020).

[41] Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 43–51.

[42] Liwen Cui, Xiaolei Xie, and Zuojun Shen. 2018. Prediction task guided representation learning of medical codes in EHR. *Journal of Biomedical Informatics* 84, (2018), 1–10. DOI: http://dx.doi.org/10.1016/j.jbi.2018.06.013

[43] Dongha Lee, Xiaoqian Jiang, and Hwanjo Yu. 2020. Harmonized representation learning on dynamic EHR graphs. *Journal of Biomedical Informatics* 106, 3 (2020), 103426. DOI: http://dx.doi.org/10.1016/j.jbi.2020.103426

[44] Siaw Teng Liaw, Jane Taggart, Hairong Yu, Simon de Lusignan, Craig Kuziemsky, and Andrew Hayen. 2014. Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers. *Journal of Biomedical Informatics* 52, (2014), 364–372. DOI: http://dx.doi.org/10.1016/j.jbi.2014.07.016

[45] Jyotishman Pathak, Sean P. Murphy, Brian N. Willaert, Hilal M. Kremers, Barbara P. Yawn, Walter A. Rocca, and Christopher G. Chute. 2011. Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: Experiences from the rochester epidemiology project. *AMIA Annual Symposium Proceedings* 2011 (2011), 1089–1098.

[46] Deborah E. Barnes, Jing Zhou, Rod L. Walker, Eric B. Larson, Sei J. Lee, W. John Boscardin, Zachary Marcum, and Sascha Dublin. 2019. Development and calidation of the electronic health rrecord risk of alzheimer's and dementia assessment rule (ERADAR). *Alzheimer's and Dementia* 68, 1(2019), 103–111. DOI: http://dx.doi.org/10.1016/j.jalz.2019.06.4579

[47] Oliver A., Chodosh J., Ferris R., and Blaum C. S. 2019. Over-treatment of older adults with diabetes and dementia. *Journal of the American Geriatrics Society* 67, S1 (2019), S120. DOI: https://doi.org/10.1111/jgs.15898

[48] Victoria Larsson, Gustav Torisson, and Elisabet Londos. 2018. Relative survival in patients with dementia with Lewy bodies and Parkinson's disease dementia. *PLoS ONE* 13, 8 (2018), e0202044. DOI: http://dx.doi.org/10.1371/journal.pone.0202044

[49] Yijun Shao, Qing T. Zeng, Kathryn K. Chen, Andrew Shutes-David, Stephen M. Thielke, and Debby W. Tsuang. 2019. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Medical Informatics and Decision Making* 19, 128 (2019), 1–11. DOI: http://dx.doi.org/10.1186/s12911-019-0846-4

[50] Joseph Bullard, Cecilia Ovesdotter Alm, Xumin Liu, Qi Yu, and Rubén Proaño. 2016. Towards early dementia detection: Fusing linguistic and non-linguistic clinical data. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*. 12–22. DOI: http://dx.doi.org/10.18653/v1/w16-0302

[51] Adam Baus PhD, MA, MPH, Jeffrey Coben MD, Keith Zullig PhD, FASHA, Cecil Pollard MA, Charles Mullett MD, PhD, Henry Taylor MD, MPH, Jill Cochran PhD, APRN-FNP, Traci Jarrett PhD, MPH, and Dustin Long PhD. 2017. An electronic health record data-driven model for identifying older adults at risk of unintentional falls. *Perspectives in Health Information Management* 14, (2017), 1–22.

[52] R. Lucero, David S. Lindberg, Elizabeth A. Fehlberg, R. Bjarnadottir, Y. Li, J. Cimiotti, Marsha Crane, and M. Prosperi. 2019. A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: Applying manual and semi- and fully-automated methods. *International Journal of Medical Informatics* 122 (2019), 63–69.

[53] S. Landis. 2013. Fall screening program in primary care practices. *Journal of the American Geriatrics Society* 62, 12 (2013), 2408–2414.

[54] James A. McCart, Donald J. Berndt, Jay Jarman, Dezon K. Finch, and Stephen L. Luther. 2013. Finding falls in ambulatory care clinical documents using statistical text mining. *Journal of the American Medical Informatics Association* 20, 5 (2013), 906–914. DOI: http://dx.doi.org/10.1136/amiajnl-2012-001334

[55] Rentsch C. T., Tate J. P., Tarko L., Honerlaw J., Cho K., Ho Y.-L., and Justice A. C. 2019. Does an index composed of routine labs discriminate risk of mortality better than the Charlson index. *Journal of General Internal Medicine* 35, (11–12) (2019), 1023–1033. DOI: http://dx.doi.org/10.1007/11606.1525-1497 LK - http://ucelinks.cdlib.org:8888/sfx_ucsf?sid=EMBASE&issn=15251497&id=doi:10.1007%2F11606.1525-1497&atitle=Does+anindex+composed+of+routine+labs+discrimi-nate+risk+of+mortality+better+than+the+charlson+index&stitle=J.+Gen.+Intern.+Med.&title=Journal+of+General+Internal+Medicine&volume=34&issue=2&spage=S194&epage=&aulast=Rentsch&aufirst=Christopher+T.&auinit=C.T.&aufull=Rentsch+C.T.&coden=&isbn=&pages=S194-&date=2019&auinit1=C&auinitm=T.

[56] Hua Xu, Melinda C. Aldrich, Qingxia Chen, Hongfang Liu, Neeraja B. Peterson, Qi Dai, Mia Levy, Anushi Shah, Xue Han, Xiaoyang Ruan, Min Jiang, Ying Li, Jamii St Julien, Jeremy Warner, Carol Friedman, Dan M. Roden, and Joshua C. Denny. 2015. Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality. *Journal of the American Medical Informatics Association* 22, 1 (2015), 179–191. DOI: http://dx.doi.org/10.1136/amiajnl-2014-002649

[57] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine* 18 (2018), 1–10. DOI: http://dx.doi.org/10.1038/s41746-018-0029-1

[58] Luchen Liu, Haoran Li, Zhiting Hu, Haoran Shi, Zichang Wang, Jian Tang, and Ming Zhang. 2019. Learning hierarchical representations of electronic health records for clinical outcome prediction. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2019. American Medical Informatics Association, 597.

[59] Jacob Calvert, Qingqing Mao, Angela J. Rogers, Christopher Barton, Melissa Jay, Thomas Desautels, Hamid Mohamadlou, Jasmine Jan, and Ritankar Das. 2016. A computational approach to mortality prediction of alcohol use disorder inpatients. *Computers in Biology and Medicine* 75, (2016), 74–79. DOI: http://dx.doi.org/10.1016/j.compbiomed.2016.05.015

[60] Matthew G. Whitbeck, Richard J. Charnigo, Paul Khairy, Khaled Ziada, Alison L. Bailey, Milagros M. Zegarra, Jignesh Shah, Gustavo Morales, Tracy MacAulay, Vincent L. Sorrell, Charles L. Campbell, John Gurley, Paul Anaya, Hafez Nasr, Rong Bai, Luigi Di Biase, David C. Booth, Guillaume Jondeau, Andrea Natale, Denis Roy, Susan Smyth, David J. Moliterno, and Claude S. Elayi. 2013. Increased mortality among patients taking digoxin - Analysis from the AFFIRM study. *European Heart Journal* 34, 20 (2013), 1481–1488. DOI: http://dx.doi.org/10.1093/eurheartj/ehs348

[61] Kevin M. Pantalone, Michael W. Kattan, Changhong Yu, Brian J. Wells, Susana Arrigain, Anil Jain, Ashish Atreja, and Robert S. Zimmerman. 2010. The risk of overall mortality in patients with type 2 diabetes receiving glipizide, glyburide, or glimepiride monotherapy: A retrospective analysis. *Diabetes Care* 33, 6 (2010), 1224–1229. DOI: http://dx.doi.org/10.2337/dc10-0017

[62] Mohammad Hashir and Rapinder Sawhney. 2020. Towards unstructured mortality prediction with free-text clinical notes. *Journal of Biomedical Informatics* 108 (2020), 103489. DOI: https://doi.org/10.1016/j.jbi.2020.103489

[63] Thomas Desautels, Jacob Calvert, Jana Hoffman, Melissa Jay, Yaniv Kerem, Lisa Shieh, David Shimabukuro, Uli Chettipally, Mitchell D. Feldman, Chris Barton, David J. Wales, and Ritankar Das. 2016. Prediction of sepsis in the intensive care unit With minimal electronic health record data: A machine learning approach. *JMIR Medical Informatics* 4, 3 (2016), e28. DOI: http://dx.doi.org/10.2196/medinform.5909

[64] Sanjeet Dadwal, Zahra Eftekhari, Tushondra Thomas, Deron Johnson, Dongyun Yang, Sally Mokhtari, Liana Nikolaenko, Janet Munu, and Ryotaro Nakamura. 2018. A machine-learning sepsis prediction model for patients undergoing hematopoietic cell transplantation. *Blood,* 132, 3 (2018), 373–374. DOI: http://dx.doi.org/10.1182/blood-2018-99-117002

[65] Ji Sun Back, Yinji Jin, Taixian Jin, and Sun Mi Lee. 2016. Development and validation of an automated sepsis risk assessment system. *Research in Nursing and Health* 39, 5 (2016), 317–327. DOI: http://dx.doi.org/10.1002/nur.21734

[66] J. Futoma, S. Hariharan, and K. Heller. 2017. Learning to detect sepsis with a multitask Gaussian process RNN classifier. (2017).

[67] Yuan Zhang, Chen Lin, Min Chi, Julie Ivy, Muge Capan, and Jeanne M. Huddleston. 2017. LSTM for septic shock: Adding unreliable labels to reliable predictions. In *Proceedings of the IEEE International Conference on Big Data, Big Data*. DOI: http://dx.doi.org/10.1109/BigData.2017.8258049

[68] Wei Qi Wei and Joshua C. Denny. 2015. Extracting Research-quality Phenotypes from Electronic Health Records to Support Precision Medicine. 7, 1 (2015), 1–14. DOI: http://dx.doi.org/10.1186/s13073-015-0166-y

[69] Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M. Biernacka, Euijung Ryu, Janet E. Olson, Mark A. Frye, Aihua Liu, Liming Guo, Ariane Marelli, Yuri Ahuja, Jose Davila-Velderrain, and Manolis Kellis. 2020. Inferring multimodal latent topics from electronic health records. *Nature Communications* 11, 1 (2020). DOI: http://dx.doi.org/10.1038/s41467-020-16378-3

[70] Michael Simmons, Ayush Singhal, and Zhiyong Lu. 2016. *Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health.* Springer Singapore, Singapore, 139–166. DOI: http://dx.doi.org/10.1007/978-981-10-1503-8_7

[71] James M. Gill, Michael S. Klinkman, and Ying Xia Chen. 2010. Antidepressant medication use for primary care patients with and without medical comorbidities: A national Electronic Health Record (EHR) network study. *Journal of the American Board of Family Medicine* 23, 4 (2010), 499–508. DOI: http://dx.doi.org/10.3122/jabfm.2010.04.090299

[72] Goodman M., Healy B., Cai T., Weiner H. L., Chitnis T., De Jager P. L., and Xia Z. 2014. Leveraging Electronic Health Records for a Phenomewide Examination of the Comorbidity Burden Associated with Multiple Sclerosis Disease Outcome. 7 (2014), e864. DOI: 10.1212/nxi.0000000000000864

[73] X. Zhang, D. K. Hayashida, and F. W. Peyerl. 2016. Analysis Of COPD comorbidities and their impact on hospital 30-day readmission rates using electronic health record data. *Value in Health* 19, 3 (2016), A110. DOI: http://dx.doi.org/10.1016/j.jval.2016.03.426

[74] Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, and Majid Sarrafzadeh. 2016. A flexible data-driven comorbidity feature extraction framework. *Computers in Biology and Medicine* 73 (2016), 165–172. DOI: http://dx.doi.org/10.1016/j.compbiomed.2016.04.014

[75] Jean-Baptiste Escudié, Bastien Rance, Georgia Malamut, Sherine Khater, Anita Burgun, Christophe Cellier, and Anne-Sophie Jannot. 2017. A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: A case study on autoimmune comorbidities in patients with celiac disease. *BMC Medical Informatics and Decision Making* 17, 1 (2017), 140.

[76] Brett K. Beaulieu-Jones and Casey S. Greene. 2016. Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics* 64 (2016), 168–178. DOI: http://dx.doi.org/10.1016/j.jbi.2016.10.007

[77] Ardavan Afshar, Ioakeim Perros, Haesun Park, Christopher deFilippi, Xiaowei Yan, Walter Stewart, Joyce Ho, and Jimeng Sun. 2020. TASTE: Temporal and static tensor factorization for phenotyping electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 193–203.

[78] Cerna A. E. U., Wehner G., Hartzel D. N., and Haggerty C. 2017. Data driven phenotyping of patients with heart failure using a deep-learning cluster representation of echocardiographic and electronic health record data. *Circulation* 136, 1 (2017).

[79] Jejo D. Koola, Sharon E. Davis, Omar Al-Nimri, Sharidan K. Parr, Daniel Fabbri, Bradley A. Malin, Samuel B. Ho, and Michael E. Matheny. 2018. Development of an automated phenotyping algorithm for hepatorenal syndrome. *Journal of Biomedical Informatics* 80 (2018), 87–95. DOI: http://dx.doi.org/10.1016/j.jbi.2018.03.001

[80] Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. 2015. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 705–714.

[81] Pedro L. Teixeira, Wei Qi Wei, Robert M. Cronin, Huan Mo, Jacob P. VanHouten, Robert J. Carroll, Eric Larose, Lisa A. Bastarache, S. Trent Rosenbloom, Todd L. Edwards, Dan M. Roden, Thomas A. Lasko, Richard A. Dart, Anne M.

Nikolai, Peggy L. Peissig, and Joshua C. Denny. 2017. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *Journal of the American Medical Informatics Association* 24, 1(2017), 162–171. DOI: http://dx.doi.org/10.1093/jamia/ocw071

[82] Benjamin S. Glicksberg, Riccardo Miotto, Kipp W. Johnson, Khader Shameer, Li Li, Rong Chen, and Joel T. Dudley. 2018. Automated disease cohort selection using word embeddings from electronic health records the creative commons attribution non-commercial (CC BY-NC) 4.0 License. HHS Public Access. *Proceedings of the Pacific Symposium on Biocomputing* 23 (2018), 145–156.

[83] Joyce C. Ho, Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. 2014. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics* 52 (2014), 199–211. DOI: http://dx.doi.org/10.1016/j.jbi.2014.07.001

[84] Wei Qi Wei, Pedro L. Teixeira, Huan Mo, Robert M. Cronin, Jeremy L. Warner, and Joshua C. Denny. 2016. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association* 23, 1(2016), 20–27. DOI: http://dx.doi.org/10.1093/jamia/ocv130

[85] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 507–516.

[86] Katherine I. Morley, Joshua Wallace, Spiros C. Denaxas, Ross J. Hunter, Riyaz S. Patel, Pablo Perel, Anoop D. Shah, Adam D. Timmis, Richard J. Schilling, and Harry Hemingway. 2014. Defining disease phenotypes using national linked electronic health records: A case study of atrial fibrillation. *PLoS ONE* 9, 11 (2014), e110900. DOI: http://dx.doi.org/10.1371/journal.pone.0110900

[87] Sheng Yu, Yumeng Ma, Jessica Gronsbell, Tianrun Cai, Ashwin N. Ananthakrishnan, Vivian S. Gainer, Susanne E. Churchill, Peter Szolovits, Shawn N. Murphy, Isaac S. Kohane, Katherine P. Liao, and Tianxi Cai. 2018. Enabling phenotypic big data with PheNorm. *Journal of the American Medical Informatics Association* 25, 1(2018), 54–60. DOI: http://dx.doi.org/10.1093/jamia/ocx111

[88] Katherine P. Liao, Ashwin N. Ananthakrishnan, Vishesh Kumar, Zongqi Xia, Andrew Cagan, Vivian S. Gainer, Sergey Goryachev, Pei Chen, Guergana K. Savova, Denis Agniel, Susanne Churchill, Jaeyoung Lee, Shawn N. Murphy, Robert M. Plenge, Peter Szolovits, Isaac Kohane, Stanley Y. Shaw, Elizabeth W. Karlson, and Tianxi Cai. 2015. Methods to develop an Electronic Medical Record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. *PLoS ONE* 10, 8 (2015), 1–11. DOI: http://dx.doi.org/10.1371/journal.pone.0136651

[89] Sheng Yu, Katherine P. Liao, Stanley Y. Shaw, Vivian S. Gainer, Susanne E. Churchill, Peter Szolovits, Shawn N. Murphy, Isaac S. Kohane, and Tianxi Cai. 2015. Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association* 22, 5 (2015), 993–1000. DOI: http://dx.doi.org/10.1093/jamia/ocv034

[90] Clayton A. Turner, Alexander D. Jacobs, Cassios K. Marques, James C. Oates, Diane L. Kamen, Paul E. Anderson, and Jihad S. Obeid. 2017. Word2Vec inversion and traditional text classifiers for phenotyping lupus. *BMC Medical Informatics and Decision Making* 22, 17(2017), 126. DOI: http://dx.doi.org/10.1186/s12911-017-0518-1

[91] Anurag Verma, Anna O. Basile, Yuki Bradford, Helena Kuivaniemi, Gerard Tromp, David Carey, Glenn S. Gerhard, James E. Crowe, Marylyn D. Ritchie, and Sarah A. Pendergrass. 2016. Phenome-Wide association study to explore relationships between immune system related genetic loci and complex traits and diseases. *PLoS ONE* 11, 8(2016), e0160573. DOI: http://dx.doi.org/10.1371/journal.pone.0160573

[92] Himanshu Sharma, Chengsheng Mao, Yizhen Zhang, Haleh Vatani, Liang Yao, Yizhen Zhong, Luke Rasmussen, Guoqian Jiang, Jyotishman Pathak, and Yuan Luo. 2019. Developing a portable natural language processing based phenotyping system. *BMC Medical Informatics and Decision Making* 19, Suppl 3 (2019). DOI: http://dx.doi.org/10.1186/s12911-019-0786-z

[93] Blosnich J. R., Montgomery A. E., Dichter M. E., Gordon A. J., Kavalieratos D., Taylor L., Ketterer B., and Bossarte R. M. 2019. Social determinants and military veterans' suicide ideation and attempt: A cross-sectional analysis of Electronic Health Record data. *Journal of General Internal Medicine* 36, 6(2019), 1759–1767. DOI: http://dx.doi.org/10.1007/s11606-019-05447-z LK - http://ucelinks.cdlib.org:8888/sfx_ucsf?sid=EMBASE&issn=15251497&id=doi:10.1007%2Fs11606-019-05447-z&atitle=Social+Determinants+and+Military+Veterans%E2%80%99+Suicide+Ideation+and+Attempt%3A+a+Cross-sectional+Analysis+of+Electronic+Health+Record+Data&stitle=J.+Gen.+Intern.+Med.&title=Journal+of+General+Internal+Medicine&volume=&issue=&spage=&epage=&aulast=Blosnich&aufirst=John+R.&auinit=J.R.&aufull=Blosnich+J.R.&coden=JGIME&isbn=&pages=-&date=2019&auinit1=J&auini.

[94] Ann John, Marcos Del Pozo Banos, and Keith Lloyd. 2018. Environmental risk factors on suicide of children and young people. *International Journal of Population Data Science* 3, 4(2018). DOI: http://dx.doi.org/10.23889/ijpds.v3i4.903

[95] Del Pozo Banos M., Lloyd K., Dennis M., Gunnel D., Scourtfield J., and John A. 2018. Case-control study of suicide in children and young people using linked primary and secondary routinely collected Electronic Health Records. *European Psychiatry* 217, 6(2018), 717–724. DOI: http://dx.doi.org/10.1016/j.eurpsy.2017.12.022 LK - http://ucelinks.cdlib.org:8888/sfx_ucsf?sid=EMBASE&issn=17783585&id=doi:10.1016%2Fj.eurpsy.2017.12.022&atitle=Case-control+study+of+suicide+in+children+and+young+people+using+linked+primary+and+secondary+routinely+collected+electronic+health+records&stitle=Eur.+Psychiatry&title=European+Psychiatry&volume=48&issue=&spage=S117&epage=&aulast=Del+Pozo+Banos&aufirst=M.&auinit=M.&aufull=Del+Pozo+Banos+M.&coden=&isbn=&pages=S117-&date=2018&auinit1=M&auinitm=.

[96] Ronald C. Kessler, Irving Hwang, Claire A. Hoffmire, John F. McCarthy, Maria V. Petukhova, Anthony J. Rosellini, Nancy A. Sampson, Alexandra L. Schneider, Paul A. Bradley, Ira R. Katz, Caitlin Thompson, and Robert M. Bossarte. 2017. Developing a practical suicide risk prediction model for targeting high-risk patients in the veterans health administration. *International Journal of Methods in Psychiatric Research* 26, 3 (2017), 1–7. DOI: http://dx.doi.org/10.1002/mpr.1575

[97] John F. McCarthy, Robert M. Bossarte, Ira R. Katz, Caitlin Thompson, Janet Kemp, Claire M. Hannemann, Christopher Nielson, and Michael Schoenbaum. 2015. Predictive modeling and concentration of the risk of suicide: Implications for preventive interventions in the US department of veterans affairs. *American Journal of Public Health* 105, 9 (2015), 1935–1942. DOI: http://dx.doi.org/10.2105/AJPH.2015.302737

[98] Jue Gong, Gregory E. Simon, and Shan Liu. 2019. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS ONE* 14, 9 (2019), e0222665. DOI: http://dx.doi.org/10.1371/journal.pone.0222665

[99] K. Haerian, H. Salmasian, and Carol Friedman. 2012. Methods for identifying suicide or suicidal ideation in EHRs. In *Proceedings of the AMIA Annual Symposium.* 1244–1253.

[100] Heather D. Anderson, Wilson D. Pace, Elias Brandt, Rodney D. Nielsen, Richard R. Allen, Anne M. Libby, David R. West, and Robert J. Valuck. 2015. Monitoring Suicidal Patients in Primary Care Using Electronic Health Records. 28, 1 (2015), 65–71. DOI: http://dx.doi.org/10.3122/jabfm.2015.01.140181

[101] Yuval Barak-Corren, Victor M. Castro, Solomon Javitt, Alison G. Hoffnagle, Yael Dai, Roy H. Perlis, Matthew K. Nock, Jordan W. Smoller, and Ben Y. Reis. 2017. Predicting suicidal behavior from longitudinal Electronic Health Records. *American Journal of Psychiatry* 174, 2 (2017), 154–162. DOI: http://dx.doi.org/10.1176/appi.ajp.2016.16010077

[102] Christopher M. Hatton, Lewis W. Paton, Dean McMillan, James Cussens, Simon Gilbody, and Paul A. Tiffin. 2019. Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders* 246, September 2018 (2019), 857–860. DOI: http://dx.doi.org/10.1016/j.jad.2018.12.095

[103] Sruthi Nelluri, Mercedes M. Rodriguez-Suarez, Zubair Rahaman, Kimberly Cabrera, Shivani Priyadarshni, Murlidhar Pamarthi, Stuti Dang, Willy Valencia, Michael J. Mintzer, and Jorge G. Ruiz. 2018. Coexisting frailty and depression in older veterans: Effects on health care utilization. *The American Journal of Geriatric Psychiatry* 35, 1(2018), 37–44. DOI: http://dx.doi.org/10.1016/j.jagp.2018.01.168

[104] Arjun Parthipan, Imon Banerjee, Keith Humphreys, Steven M. Asch, Catherine Curtin, Ian Carroll, and Tina Hernandez-Boussard. 2019. Predicting inadequate postoperative pain management in depressed patients: A machine learning approach. *PLoS ONE* 14, 2 (2019), e0210575. DOI: http://dx.doi.org/10.1371/journal.pone.0210575

[105] Sandy H. Huang, Paea LePendu, Srinivasan V. Iyer, Ming Tai-Seale, David Carrell, and Nigam H. Shah. 2014. Toward personalizing treatment for depression: Predicting diagnosis and severity. *Journal of the American Medical Informatics Association* 21, 6 (2014), 1069–1075. DOI: http://dx.doi.org/10.1136/amiajnl-2014-002733

[106] Wenshuo Liu, Cooper Stansbury, Karandeep Singh, Andrew M. Ryan, Devraj Sukul, Elham Mahmoudi, Akbar Waljee, Ji Zhu, and Brahmajee K. Nallamothu. 2020. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PloS ONE* 15, 4 (2020), e0221606.

[107] Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Sławomir Nowaczyk. 2019. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics* 97 (2019), 103256.

[108] K. Noon, N. Sarabu, J. Augustine, D. Hricik, B. Deleva, K. Woodside, M. Aeder, J. Foote, A. Bruno, K. Walsh, M. Johnston, V. Humphreville, and E. Sanchez. 2016. Effect of telehealth monitoring on early hospital readmission after renal transplantation. In *Proceedings of the American Journal of Transplantation.* 684.

[109] Juliessa M. Pavon, Yanfang Zhao, Eleanor McConnell, and S. Nicole Hastings. 2014. Identifying risk of readmission in hospitalized elderly adults through inpatient medication exposure. *Journal of the American Geriatrics Society* 62, 6 (2014), 1116–1121. DOI: http://dx.doi.org/10.1111/jgs.12829

[110] Navathe A. S., Zhong F., Lei V., Chang F. Y., Rocha R. A., and Zhou L. 2016. Improving identification of patients at high-risk for readmission using socio-behavioral patient characteristics. *Journal of General Internal Medicine* 13, 6(2016), 1070–1081.

[111] Jeffrey L. Greenwald, Patrick R. Cronin, Victoria Carballo, Goodarz Danaei, and Garry Choy. 2017. A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial

support using natural language processing. *Medical Care* 55, 3 (2017), 261–266. DOI: http://dx.doi.org/10.1097/MLR.0000000000000651

[112] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv:1904.05342. Retrieved from http://arxiv.org/abs/1904.05342.

[113] Xi Sheryl Zhang, Fengyi Tang, Hiroko H. Dodge, Jiayu Zhou, and Fei Wang. 2019. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2487–2495.

[114] Stacey E. Jolly, Sankar D. Navaneethan, Jesse D. Schold, Susana Arrigain, John W. Sharp, Anil K. Jain, Martin J. Schreiber, James F. Simon, and Joseph V. Nally. 2014. Chronic kidney disease in an Electronic Health Record problem list: Quality of care, ESRD, and mortality. *American Journal of Nephrology* 39, 4(2014), 288–96. DOI: http://dx.doi.org/10.1159/000360306

[115] Michael Simonov, Ugochukwu Ugwuowo, Erica Moreira, Yu Yamamoto, Aditya Biswas, Melissa Martin, Jeffrey Testani, and F. Perry Wilson. 2019. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: A descriptive modeling study. *PLoS Medicine* 16, 7 (2019), e1002861. DOI: http://dx.doi.org/10.1371/journal.pmed.1002861

[116] Samuel J. Weisenthal, Caroline Quill, Samir Farooq, Henry Kautz, and Martin S. Zand. 2018. Predicting acute kidney injury at hospital reentry using high-dimensional Electronic Health Record data. *PLoS ONE* 13, 11 (2018), e0204920. DOI: http://dx.doi.org/10.1371/journal.pone.0204920

[117] Sankar D. Navaneethan, Stacey E. Jolly, Jesse D. Schold, Susana Arrigain, Welf Saupe, John Sharp, Jennifer Lyons, James F. Simon, Martin J. Schreiber, Anil Jain, and Joseph V. Nally. 2011. Development and validation of an Electronic Health Record-based chronic kidney disease registry. *Clinical Journal of the American Society of Nephrology* 6, 1(2011), 40–49. DOI: http://dx.doi.org/10.2215/CJN.04230510

[118] Michael E. Matheny, Randolph A. Miller, T. Alp Ikizler, Lemuel R. Waitman, Joshua C. Denny, Jonathan S. Schildcrout, Robert S. Dittus, and Josh F. Peterson. 2010. Development of inpatient risk stratification models of acute kidney injury for use in Electronic Health Records. *Medical Decision Making* 30, 6(2010), 639–650. DOI: http://dx.doi.org/10.1177/0272989X10364246

[119] Michele Bernardini, Luca Romeo, Paolo Misericordia, and Emanuele Frontoni. 2020. Discovering the type 2 Diabetes in Electronic Health Records using the sparse balanced support vector machine. *IEEE Journal of Biomedical and Health Informatics* 24, 1(2020), 235–246. DOI: http://dx.doi.org/10.1109/JBHI.2019.2899218

[120] Hang Qiu, Hai Yan Yu, Li Ya Wang, Qiang Yao, Si Nan Wu, Can Yin, Bo Fu, Xiao Juan Zhu, Yan Long Zhang, Yong Xing, Jun Deng, Hao Yang, and Shun Dong Lei. 2017. Electronic Health Record driven prediction for gestational diabetes mellitus in early pregnancy. *Scientific Reports* 7, 1(2017), 16417. DOI: http://dx.doi.org/10.1038/s41598-017-16665-y

[121] Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. 2017. A machine learning-based framework to identify type 2 diabetes through Electronic Health Records. *International Journal of Medical Informatics* 97 (2017), 120–127. DOI: http://dx.doi.org/10.1016/j.ijmedinf.2016.09.014

[122] Victor W. Zhong, Jihad S. Obeid, Jean B. Craig, Emily R. Pfaff, Joan Thomas, Lindsay M. Jaacks, Daniel P. Beavers, Timothy S. Carey, Jean M. Lawrence, Dana Dabelea, Richard F. Hamman, Deborah A. Bowlby, Catherine Pihoker, Sharon H. Saydah, and Elizabeth J. Mayer-Davis. 2016. An efficient approach for surveillance of childhood diabetes by type derived from Electronic Health Record data: The SEARCH for diabetes in youth study. *Journal of the American Medical Informatics Association* 23, 6(2016), 1060–1067. DOI: http://dx.doi.org/10.1093/jamia/ocv207

[123] Michael Klompas, Emma Eggleston, Jason McVetta, Ross Lazarus, Lingling Li, and Richard Platt. 2013. Automated detection and classification of type 1 versus type 2 diabetes using Electronic Health Record data. *Diabetes Care* 36, 4 (2013), 914–921. DOI: http://dx.doi.org/10.2337/dc12-0964

[124] J. P. Anderson, J. R. Parikh, D. K. Shenfeld, B. W. Church, J. M. Laramie, B. A. Piper, R. J. Willke, J. Mardekian, and D. A. Rublee. 2014. Identification of determinants of progression to type 2 diabetes using electronic health records and 'big data' analytics. *Value in Health* 17, 3(2014), A242. DOI: http://dx.doi.org/10.1016/j.jval.2014.03.1413

[125] Adam Wright, Allison B. McCoy, Stanislav Henkin, Abhivyakti Kale, and Dean F. Sittig. 2013. Use of a support vector machine for categorizing free-text notes: Assessment of accuracy across two institutions. *Journal of the American Medical Informatics Association* 20, 5 (2013), 887–980. DOI: http://dx.doi.org/10.1136/amiajnl-2012-001576

[126] G. Maragatham and Shobana Devi. 2019. LSTM model for prediction of heart failure in big data. *Journal of Medical Systems* 43, 5 (2019), 111. DOI: http://dx.doi.org/10.1007/s10916-019-1243-3

[127] Rui Zhang, Sisi Ma, Liesa Shanahan, Jessica Munroe, Sarah Horn, and Stuart Speedie. 2018. Discovering and identifying New York heart association classification from Electronic Health Records. *BMC Medical Informatics and Decision Making* 18, 2 (2018), 48. DOI: http://dx.doi.org/10.1186/s12911-018-0625-7

[128] Maryam Panahiazar, Vahid Taslimitehrani, Naveen L. Pereira, and Jyotishman Pathak. 2015. Using EHRs for heart failure therapy recommendation using multidimensional patient similarity aSnalytics. In *Studies in Health Technology and Informatics*, Vol. 210. 369. DOI: http://dx.doi.org/10.3233/978-1-61499-512-8-369

[129] Edward Choi, Cao Xiao, Jimeng Sun, and Walter F. Stewart. 2018. Mime: Multilevel medical embedding of Electronic Health Records for predictive healthcare. In *Proceedings of the Advances in Neural Information Processing Systems.*

[130] Roy J. Byrd, Steven R. Steinhubl, Jimeng Sun, Shahram Ebadollahi, and Walter F. Stewart. 2014. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from Electronic Health Records. *International Journal of Medical Informatics* 83, 12(2014), 983–992. DOI: http://dx.doi.org/10.1016/j.ijmedinf.2012.12.005

[131] Mark E. Patterson, Derick Miranda, Gregory L. Schuman, Christopher M. Eaton, Andrew J. Smith, and Brad Silver. 2016. A focus group exploration of automated case-finders to identify high-risk heart failure patients within an urban safety-net hospital. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)* 4, 3(2016), 1225. DOI: http://dx.doi.org/10.13063/2327-9214.1225

[132] Yirong Wu, Elizabeth S. Burnside, Jennifer Cox, Jun Fan, Ming Yuan, Jie Yin, Peggy Peissig, Alexander Cobian, David Page, and Mark Craven. 2017. Breast cancer risk prediction using Electronic Health Records. In *Proceedings of the IEEE International Conference on Healthcare Informatics, ICHI.* DOI: http://dx.doi.org/10.1109/ICHI.2017.62

[133] Di Zhao and Chunhua Weng. 2011. Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of Biomedical Informatics* 44, 5 (2011), 859–868. DOI: http://dx.doi.org/10.1016/j.jbi.2011.05.004

[134] Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B. Peterson, Qingxia Chen, Subramani Mani, Mia A. Levy, Q. Dai, and Josh C. Denny. 2011. Extracting and integrating data from entire Electronic Health Records for detecting colorectal cancer cases. In *Proceedings of the Annual AMIA Symposium* (2011).

[135] Alexander W. Forsyth, Regina Barzilay, Kevin S. Hughes, Dickson Lui, Karl A. Lorenz, Andrea Enzinger, James A. Tulsky, and Charlotta Lindvall. 2018. Machine learning methods to extract documentation of breast cancer symptoms from Electronic Health Records. *Journal of Pain and Symptom Management* 55, 6 (2018), 1492–1499. DOI: http://dx.doi.org/10.1016/j.jpainsymman.2018.02.016

[136] A. Walling, S. F. D'Ambruoso, S. Hurvitz, R. Clarke, A. Hackbarth, C. Pietras, and N. Wenger. 2015. A palliative nurse practitioner intervention to improve advance care planning and supportive care in patients with advanced cancer. *Journal of General Internal Medicine* (2015), S87–S87.

[137] Po Yen Wu, Chih Wen Cheng, Chanchala D. Kaddi, Janani Venugopalan, Ryan Hoffman, and May D. Wang. 2017. -Omic and Electronic Health Record dig data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering* 64, 2 (2017), 263–273. DOI: http://dx.doi.org/10.1109/TBME.2016.2573285

[138] Max Robinson, Jennifer Hadlock, Jiyang Yu, Alireza Khatamian, Aleksandr Y. Aravkin, Eric W. Deutsch, Nathan D. Price, Sui Huang, and Gustavo Glusman. 2018. Fast and simple comparison of semi-structured data, with emphasis on Electronic Health Records. *bioRxiv* (2018). DOI: http://dx.doi.org/10.1101/293183

[139] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (2010), 507–513.

[140] Qi Li, Stephen Spooner, Megan Kaiser, Nataline Lingren, Jessica Robbins, Todd Lingren, Huaxiu Tang, Imre Solti, and Yizhao Ni. 2015. An end-to-end hybrid algorithm for automated medication discrepancy detection. *BMC Medical Informatics and Decision Making* 15 (05 2015), 37. DOI: http://dx.doi.org/10.1186/s12911-015-0160-8

[141] C. Zheng, N. Rashid, River Koblick, and J. An. 2015. Medication extraction from electronic clinical notes in an integrated health system: A study on aspirin use in patients with nonvalvular atrial fibrillation. *Clinical Therapeutics* 37, 9 (2015), 2048–2058.e2.

[142] Licong Cui, Alireza Bozorgi, Samden Lhatoo, Guo-Qiang Zhang, and Satya Sahoo. 2012. EpiDEA: Extracting structured Epilepsy and Seizure information from patient discharge summaries for cohort identification. *AMIA Annual Symposium Proceedings* 2012 (11 2012), 1191–200.

[143] Son Doan, Mike Conway, Tu Phuong, and Lucila Ohno-Machado. 2014. Natural language processing in biomedicine: A unified system architecture overview. *Clinical Bioinformatics* 1168 (2014), 275–294.

[144] Saeed Hassanpour and Curtis Langlotz. 2015. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine* 66 (10 2015). DOI: http://dx.doi.org/10.1016/j.artmed.2015.09.007

[145] Tianrun Cai, Andreas Giannopoulos, Sheng Yu, Tatiana Kelil, Beth Ripley, Kanako Kumamaru, Frank Rybicki, and Dimitrios Mitsouras. 2016. Natural language processing technologies in radiology research and clinical applications. *Radiographics: A Review Publication of the Radiological Society of North America, Inc* 36 (01 2016), 176–191. DOI: http://dx.doi.org/10.1148/rg.2016150080

[146] Chen Lin, Elizabeth W. Karlson, Dmitriy Dligach, Monica P. Ramirez, Timothy A. Miller, Huan Mo, Natalie S. Braggs, Andrew Cagan, Vivian Gainer, Joshua C. Denny, and Guergana K. Savova. 2014. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association* 22, e1 (10 2014), e151–e161. DOI: http://dx.doi.org/10.1136/amiajnl-2014-002642 arXiv:https://academic.oup.com/jamia/article-pdf/22/e1/e151/34146153/amiajnl-2014-002642.pdf.

[147] Vijay Garla, III Re, Vincent Lo, Zachariah Dorey-Stein, Farah Kidwai, Matthew Scotch, Julie Womack, Amy Justice, and Cynthia Brandt. 2011. The Yale cTAKES extensions for document classification: Architecture and application. *Journal of the American Medical Informatics Association* 18, 5 (05 2011), 614–620. DOI: http://dx.doi.org/10.1136/amiajnl-2011-000093 arXiv:https://academic.oup.com/jamia/article-pdf/18/5/614/5965075/18-5-641.pdf.

[148] David Martinez, Graham Pitson, Andrew MacKinlay, and Lawrence Cavedon. 2014. Cross-hospital portability of information extraction of cancer staging information. *Artificial Intelligence in Medicine* 62, 1 (2014), 11–21. DOI: http://dx.doi.org/10.1016/j.artmed.2014.06.002

[149] Naveen Ashish, Lisa Dahm, and Charles Boicey. 2014. University of California, Irvine-pathology extraction pipeline: The pathology extraction pipeline for information extraction from pathology reports. *Health Informatics Journal* 20, 4 (2014), 288–305. DOI: http://dx.doi.org/10.1177/1460458213494032

[150] Kabir Yadav, Efsun Sarioglu, Hyeong Choi, Walter Cartwright, Pamela Hinds, and James Chamberlain. 2016. Automated Outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine* 23, 2 (01 2016), 171–178. DOI: http://dx.doi.org/10.1111/acem.12859

[151] Shumei Yin, Chunying Li, Yigang Zhou, and Jun Huang. 2013. Detecting hotspots in Insulin-like growth factors 1 research through MetaMap and data mining technologies. In *Proceedings of the International Conference on Web Information Systems Engineering Proceedings*, Zhisheng Huang, Chengfei Liu, Jing He, and Guangyan Huang (Eds.), Springer Berlin Heidelberg, Berlin, 359–372.

[152] Ioannis Korkontzelos, Dimitrios Piliouras, Andrew W. Dowsey, and Sophia Ananiadou. 2015. Boosting drug named entity recognition using an aggregate classifier. *Artificial Intelligence in Medicine* 65, 2 (2015), 145–153. DOI: http://dx.doi.org/10.1016/j.artmed.2015.05.007 Intelligent healthcare informatics in big data era.

[153] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations.* 1–12. DOI: http://dx.doi.org/10.1162/153244303322533223

[154] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the NAACL HLT 2018 - Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference.* DOI: http://dx.doi.org/10.18653/v1/n18-1202

[155] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 abs/1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

[156] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, 1471–4186. DOI: 10.18653/v1/N19-1423

[157] Mary Jane C. Samonte, Bobby D. Gerardo, Arnel C. Fajardo, and Ruji P. Medina. 2018. ICD-9 tagging of clinical notes using topical word embedding. In *Proceedings of the* 2018 *International Conference on Internet and e-Business.* 118–123. DOI: http://dx.doi.org/10.1145/3230348.3230357

[158] Zhuoran Wang, Anoop D. Shah, A. Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, and Harry Hemingway. 2012. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 7, 1 (2012), e30412.

[159] Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2019. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Proceedings of the Pacific Symposium on Biocomputing.* World Scientific. 295–306.

[160] Olivier Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 32, suppl_1 (2004), D267–D270.

[161] Elyne Scheurwegs, Boris Cule, Kim Luyckx, Léon Luyten, and Walter Daelemans. 2017. Selecting relevant features from the Electronic Health Record for clinical code prediction. *Journal of Biomedical Informatics* 74 (2017), 92–110. DOI: http://dx.doi.org/10.1016/j.jbi.2017.09.004

[162] Jason Scott Mathias, Ankit Agrawal, Joe Feinglass, Andrew J. Cooper, David William Baker, and Alok Choudhary. 2013. Development of a 5 year life expectancy index in older adults using predictive mining of Electronic Health Record data. *Journal of the American Medical Informatics Association* 20, 1 (2013), 118–124. DOI: http://dx.doi.org/10.1136/amiajnl-2012-001360

[163] Ravi Garg, Elissa Oh, Andrew Naidech, Konrad Kording, and Shyam Prabhakaran. 2019. Automating ischemic stroke subtype classification using machine learning and natural language processing. *Journal of Stroke and Cerebrovascular Diseases* 28, 7 (2019), 2045–2051. DOI: http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2019.02.004

[164] Yanshan Wang, Yiqing Zhao, Terry M. Therneau, Elizabeth J. Atkinson, Ahmad P. Tafti, Nan Zhang, Shreyasee Amin, Andrew H. Limper, Sundeep Khosla, and Hongfang Liu. 2020. Unsupervised machine learning for the discovery of

latent disease clusters and patient subgroups using electronic health records. *Journal of Biomedical Informatics* 102 (2020), 103364. DOI: http://dx.doi.org/10.1016/j.jbi.2019.103364 arXiv:1905.10309

[165] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. 2016. Deep patient: An unsupervised representation to predict the future of patients from the Electronic Health Records. *Scientific Reports* 6 (2016), 1–10. DOI: http://dx.doi.org/10.1038/srep26094

[166] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1903–1911.

[167] Patrick Wu, Aliya Gifford, Xiangrui Meng, Xue Li, Harry Campbell, Tim Varley, Juan Zhao, Robert Carroll, Lisa Bastarache, Joshua C. Denny, Evropi Theodoratou, and Wei-Qi Wei. 2019. Developing and evaluating mappings of ICD-10 and ICD-10-CM codes to phecodes. *JMIR Medical Informatics* 7, 4 (2019), e14325. DOI: http://dx.doi.org/10.2196/14325

[168] Hadi Kharrazi, Laura J. Anzaldi, Leilani Hernandez, Ashwini Davison, Cynthia M. Boyd, Bruce Leff, Joe Kimura, and Jonathan P. Weiner. 2018. The value of unstructured Electronic Health Record data in geriatric syndrome case identification. *Journal of the American Geriatrics Society* 66, 8(2018), 1499–1507. DOI: http://dx.doi.org/10.1111/jgs.15411

[169] Sharona Hoffman and Andy Podgurski. 2013. Big bad data: Law, public health, and biomedical databases. *The Journal of Law, Medicine & Ethics* 41 (2013), 56–60.

[170] Alison Callahan, Nigam H. Shah, and Jonathan H. Chen. 2020. Research and reporting considerations for observational studies using Electronic Health Record data. *Annals of Internal Medicine* 172, 11_Supplement (2020), S79–S84.

[171] George Hripcsak and David J. Albers. 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 20, 1 (2013), 117–121. DOI: http://dx.doi.org/10.1136/amiajnl-2012-001145

[172] Erikson Júlio De Aguiar, Bruno S. Faiçal, Bhaskar Krishnamachari, and Jó Ueyama. 2020. A survey of blockchain-based strategies for healthcare. *Computing Surveys* 53, 2 (2020), 1–27.

[173] Thomas Ploug and Søren Holm. 2020. The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care and Philosophy* 23, 1 (2020), 107–114.

[174] Lisa M. Lee. 2017. Ethics and subsequent use of Electronic Health Record data. *Journal of Biomedical Informatics* 71 (2017), 143–146.

[175] E. Andrew Balas, Marlo Vernon, Farah Magrabi, Lynne Thomas Gordon, Joanne Sexton, et al. 2015. Big data clinical research: Validity, ethics, and regulation. In *Proceedings of the MedInfo.* 448–452.

[176] Daniel Schönberger. 2019. Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* 27, 2 (2019), 171–203.

[177] Nicole Martinez-Martin, Thomas R. Insel, Paul Dagum, Henry T. Greely, and Mildred K. Cho. 2018. Data mining for health: Staking out the ethical territory of digital phenotyping. *NPJ Digital Medicine* 1, 1 (2018), 1–5. DOI: http://dx.doi.org/10.1038/s41746-018-0075-8

[178] Tamra Lysaght, Hannah Yeefen Lim, Vicki Xafis, and Kee Yuan Ngiam. 2019. AI-assisted decision-making in healthcare. *Asian Bioethics Review* 11, 3 (2019), 299–314. DOI: http://dx.doi.org/10.1007/s41649-019-00096-0

[179] Tim Jacquemard, Colin P. Doherty, and Mary B. Fitzsimons. 2020. Examination and diagnosis of electronic patient records and their associated ethics: A scoping literature review. *BMC Medical Ethics* 21, 1 (2020), 1–13.

[180] James L. Bernat. 2013. Ethical and quality pitfalls in Electronic Health Records. *Neurology* 80, 11 (2013), 1057–1061.

[181] Sjoukje van der Bij, Nasra Khan, Petra Ten Veen, Dinny H. de Bakker, and Robert A. Verheij. 2017. Improving the quality of EHR recording in primary care: A data quality feedback tool. *Journal of the American Medical Informatics Association* 24, 1 (2017), 81–87.

[182] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. 2010. Secondary use of EHR: Data quality issues and informatics opportunities. *Summit on Translational Bioinformatics* 2010 (2010), 1.

[183] Sebastien Haneuse. 2016. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical Care* 54, 4 (2016), e23.

[184] Benjamin A. Goldstein, Nrupen A. Bhavsar, Matthew Phelan, and Michael J. Pencina. 2016. Controlling for informed presence bias due to the number of health encounters in an Electronic Health Record. *American Journal of Epidemiology* 184, 11 (2016), 847–855.

[185] Andrea C. Skelly, Joseph R. Dettori, and Erika D. Brodt. 2012. Assessing bias: The importance of considering confounding. *Evidence-Based Spine-Care Journal* 3, 1 (2012), 9.

[186] Julie K. Bower, Sejal Patel, Joyce E. Rudy, and Ashley S. Felix. 2017. Addressing bias in Electronic Health Record-based surveillance of cardiovascular disease risk: Finding the signal through the noise. *Current Epidemiology Reports* 4, 4 (2017), 346–352.

[187] Matthew Phelan, Nrupen A. Bhavsar, and Benjamin A. Goldstein. 2017. Illustrating informed presence bias in Electronic Health Records data: How patient interactions with a health system can impact inference. *eGEMs(Generating Evidence & Methods to Improve Patient Outcomes)* 5, 1 (2017), 22.

[188] Giovanni Tripepi, Kitty J. Jager, Friedo W. Dekker, and Carmine Zoccali. 2010. Selection bias and information bias in clinical research. *Nephron Clinical Practice* 115, 2 (2010), c94–c99.

[189] Christopher W. Snyder, Jordan A. Weinberg, Gerald McGwin Jr, Sherry M. Melton, Richard L. George, Donald A. Reiff, James M. Cross, Jennifer Hubbard-Brown, Loring W. Rue III, and Jeffrey D. Kerby. 2009. The relationship of blood product ratio to mortality: Survival benefit or survival bias? *Journal of Trauma and Acute Care Surgery* 66, 2 (2009), 358–364.

[190] Ruth Farmer, Rohini Mathur, Krishnan Bhaskaran, Sophie V. Eastwood, Nish Chaturvedi, and Liam Smeeth. 2018. Promises and pitfalls of Electronic Health Record analysis. *Diabetologia* 61, 6 (2018), 1241–1248.

[191] Gaurav Jetley and He Zhang. 2019. Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. *Decision Support Systems* 126 (2019), 113137.

[192] Jessica S. Ancker, Sarah Shih, Mytri P. Singh, Andrew Snyder, Alison Edwards, and Rainu Kaushal. 2011. Root causes underlying challenges to secondary use of data. In *Proceedings of the AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 57–.

[193] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. 2013. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS ONE* 8, 6 (2013), e66341.

[194] Kristin M. Corey, Sehj Kashyap, Elizabeth Lorenzi, Sandhya A. Lagoo-Deenadayalan, Katherine Heller, Krista Whalen, Suresh Balu, Mitchell T. Heflin, Shelley R. McDonald, Madhav Swaminathan, et al. 2018. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated Electronic Health Record data (Pythia): A retrospective, single-site study. *PLoS Medicine* 15, 11 (2018), e1002701.

[195] Zongyin Peng, Guiling Xu, Hui Zhou, Yu Yao, Hui Ren, Jieling Zhu, Hui Liu, and Wen Liu. 2019. Early warning of nursing risk based on patient electronic medical record information. *Journal of Infection and Public Health* 13, 10 (2019), 1562–1566.

[196] Weicheng Zhu and Narges Razavian. 2019. Graph neural network on Electronic Health Records for predicting Alzheimer's Disease. *ArXiv* arXiv:1912.03761. Retrieved from https://arxiv.org/abs/1912.03761.

[197] Zachary Chase Lipton, David C. Kale, Charles Elkan, and Randall C. Wetzel. 2016. Learning to diagnose with LSTM recurrent neural networks. In *Proceedings of 4th International Conference on Learning Representations, ICLR*, Yoshua Bengio and Yann LeCun (Eds.), Retrieved from http://arxiv.org/abs/1511.03677.

[198] Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific Data* 6, 1 (2019), 1–18. DOI:http://dx.doi.org/10.1038/s41597-019-0103-9

[199] Saaed Mehrabi, Sunghwan Sohn, Dingheng Li, Joshua J. Pankratz, Terry Therneau, Jennifer L St Sauver, Hongfang Liu, and Mathew Palakal. 2015. Temporal pattern and association discovery of diagnosis codes using deep learning. In *Proceedings of the International Conference on Healthcare Informatics*. IEEE, 408–416.

[200] Wen Wang, Han Zhao, Honglei Zhuang, Nirav Shah, and Rema Padman. 2020. DyCRS: Dynamic interpretable post-operative complication risk scoring. In *Proceedings of The Web Conference*. 1839–1850.

[201] Chen Zhan, Elizabeth Roughead, Lin Liu, Nicole Pratt, and Jiuyong Li. 2020. Detecting potential signals of adverse drug events from prescription data. *Artificial Intelligence in Medicine* 104 (2020), 101839.

[202] Yang Xiang, Jun Xu, Yuqi Si, Zhiheng Li, Laila Rasmy, Yujia Zhou, Firat Tiryaki, Fang Li, Y. Zhang, Y. Wu, Xiaoqian Jiang, W. J. Zheng, D. Zhi, Cui Tao, and Wang Qi. 2019. Time-sensitive clinical concept embeddings learned from large electronic health records. *BMC Medical Informatics and Decision Making* 19 (2019), 139–148.

[203] Chao Che, Cao Xiao, Jian Liang, Bo Jin, Jiayu Zho, and Fei Wang. 2017. An RNN architecture with dynamic temporal matching for personalized predictions of Parkinson's disease. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 198–206.

[204] Hye Jin Kam and Ha Young Kim. 2017. Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine* 89 (2017), 248–255.

[205] Fei Wang, Jiayu Zhou, and Jianying Hu. 2014. DensityTransfer: A data driven approach for imputing Electronic Health Records. In *Proceedings of the 22nd International Conference on Pattern Recognition*. IEEE, 2763–2768.

[206] Francesco Bagattini, Isak Karlsson, Jonathan Rebane, and Panagiotis Papapetrou. 2019. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Medical Informatics and Decision Making* 19, 1 (2019), 7.

[207] Xi Yang, Yuan Zhang, and Min Chi. 2018. Time-aware subgroup matrix decomposition: Imputing missing data using forecasting events. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*. IEEE, 1524–1533.

[208] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 65–74.

[209] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* 8, 1 (2018), 1–12.

[210] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M. Glass, and Jimeng Sun. 2020. StageNet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference*. 530–540.

[211] Mehak Gupta, Thao-Ly T. Phan, Timothy Bunnell, and Rahmatollah Beheshti. 2019. Obesity prediction with EHR Data: A deep learning approach with interpretable elements. arXiv: Applications (2019), arXiv–1912.

[212] Qingxiong Tan, Mang Ye, Baoyao Yang, Siqi Liu, Andy Jinhua Ma, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and PongChi Yuen. 2020. DATA-GRU: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 930–937.

[213] Yeo-Jin Kim and Min Chi. 2018. Temporal belief memory: Imputing missing data during RNN training. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI).*

[214] Mohamed Ghalwash, Ying Li, Ping Zhang, and Jianying Hu. 2017. Exploiting electronic health records to mine drug effects on laboratory test results. In *Proceedings of ACM Conference on Information and Knowledge Management*. 1837–1846.

[215] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics* 69 (2017), 218–229.

[216] Kaiping Zheng, Wei Wang, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and Wei Luen James Yip. 2017. Capturing feature-level irregularity in disease progression modeling. In *Proceedings of ACM Conference on Information and Knowledge Management*. 1579–1588.

[217] Andreas Storvik Strauman, Filippo Maria Bianchi, Karl Øyvind Mikalsen, Michael Kampffmeyer, Cristina Soguero-Ruiz, and Robert Jenssen. 2018. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks. In *Proceedings of the 2018 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 307–310.

[218] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 30–41.

[219] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. 2017. Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics* 21, 1 (2017), 22–30.

[220] Tao Chen, Mark Dredze, Jonathan P. Weiner, Leilani Hernandez, Joe Kimura, and Hadi Kharrazi. 2019. Extraction of geriatric syndromes from Electronic Health Record clinical notes: Assessment of statistical natural language processing methods. *JMIR Medical Informatics* 7, 1 (2019), e13039.

[221] Tao Chen, Mark Dredze, Jonathan P. Weiner, and Hadi Kharrazi. 2019. Identifying vulnerable older adult populations by contextualizing geriatric syndrome information in clinical notes of electronic health records. *Journal of the American Medical Informatics Association* 26, 8–9 (2019), 787–795.

[222] Richard G. Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J. Dobson, and Robert Stewart. 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* 7, 1 (2017).

[223] Min Jiang, Yukun Chen, Mei Liu, S. Trent Rosenbloom, Subramani Mani, Joshua C. Denny, and Hua Xu. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association* 18, 5 (2011), 601–606.

[224] Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B. Ellis, Erwin P. Bottinger, and John V. Guttag. 2015. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *Journal of Biomedical Informatics* 53 (2015), 220–228.

[225] Anoop D. Shah, Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology* 179, 6 (2014), 764–774.

[226] Tengfei Ma, Cao Xiao, and Fei Wang. 2018. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 261–269.

[227] Jinghe Zhang, Jiaqi Gong, and Laura Barnes. 2017. HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In *Proceedings of the IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*. IEEE, 214–221.

[228] George Hripcsak, David J. Albers, and Adler Perotte. 2015. Parameterizing time in Electronic Health Record studies. *Journal of the American Medical Informatics Association* 22, 4 (2015), 794–804.

[229] Brett K. Beaulieu-Jones, Daniel R. Lavage, John W. Snyder, Jason H. Moore, Sarah A. Pendergrass, and Christopher R. Bauer. 2018. Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR Medical Informatics* 6, 1 (2018), e11.

[230] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 85–94.

[231] Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* 147, 1 (1981), 195–197.

[232] Ying Sha, Janani Venugopalan, and May D. Wang. 2016. A novel temporal similarity measure for patients based on irregularly measured data in electronic health records. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 337–344.

[233] Matthew Sperrin, Sarah Thew, James Weatherall, William Dixon, and Iain Buchan. 2011. Quantifying the longitudinal value of healthcare record collections for pharmacoepidemiology. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 1318.

[234] Kai He, Shuai Huang, and Xiaoning Qian. 2019. Early detection and risk assessment for chronic disease with irregular longitudinal data analysis. *Journal of Biomedical Informatics* 96 (2019), 103231.

[235] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BeHRt: Transformer for Electronic Health Records. *Scientific Reports* 10, 1 (2020), 1–12.

[236] Uiwon Hwang, Sungwoon Choi, Han-Byoel Lee, and Sungroh Yoon. 2017. Adversarial training for disease prediction from electronic health records with missing data. arXiv:1711.04126. Retrieved from https://arxiv.org/abs/1711.04126.

[237] Zachary C. Lipton, David Kale, and Randall Wetzel. 2016. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research)*, Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens (Eds.), Vol. 56. PMLR, Children's Hospital LA, Los Angeles, CA, 253–270. Retrieved from http://proceedings.mlr.press/v56/Lipton16.html.

[238] Ran Ilan Ber and Tom Haramaty. 2020. Domain adaptation in highly imbalanced and overlapping datasets. arXiv:2005.03585. Retrieved from https://arxiv.org/abs/2005.03585.

[239] Yang Zhao, Zoie Shui-Yee Wong, and Kwok Leung Tsui. 2018. A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering* 2018 (2018), 1–11.

[240] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Aziz Sheikh, and Majda Pajnkihar. 2019. Challenges associated with missing data in Electronic Health Records: A case study of a risk prediction model for diabetes using data from Slovenian primary care. *Health Informatics Journal* 25, 3 (2019), 951–959.

[241] Cristina Soguero-Ruiz, Wang M. E. Fei, Robert Jenssen, Knut Magne Augestad, José-Luis Rojo Álvarez, Inmaculada Mora Jiménez, Rolv-Ole Lindsetmo, and Stein Olav Skrøvseth. 2015. Data-driven temporal prediction of surgical site infection. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2015. American Medical Informatics Association, 1164.

[242] Aaron N. Richter and Taghi M. Khoshgoftaar. 2018. Building and interpreting risk models from imbalanced clinical data. In *Proceedings of the IEEE 30th International Conference on Tools with Artificial Intelligence*. IEEE, 143–150.

[243] Karmen S. Williams and Gulzar H. Shah. 2016. Electronic Health Records and meaningful use in local health departments: Updates from the 2015 NACCHO Informatics Assessment Survey. *Journal of Public Health Management and Practice* 22, Suppl 6 (2016), S27.

[244] Emran Saleh, Aïda Valls, Antonio Moreno, Pedro Romero-Aroca, and Sanitaria Pere Virgili. 2017. Integration of different fuzzy rule-induction methods to improve the classification of patients with diabetic retinopathy. In *Proceedings of the CCIA*. 6–15.

[245] Chris Drummond, Robert C. Holte, et al. 2003. C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the Workshop on Learning from Imbalanced Datasets*. 11. Citeseer, 1–8.

[246] Robert C. Holte, Liane E. Acker, and Bruce W. Porter. 1989. Concept learning and the problem of small disjuncts. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1* . Morgan Kaufmann Publishers Inc., San Francisco, CA, 813–818.

[247] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[248] Binh P. Nguyen, Hung N. Pham, Hop Tran, Nhung Nghiem, Quang H. Nguyen, Trang T. T. Do, Cao Truong Tran, and Colin R. Simpson. 2019. Predicting the onset of type 2 diabetes using wide and deep learning with Electronic Health Records. *Computer Methods and Programs in Biomedicine* 182 (2019), 105055.

[249] Andrew Maxwell, Runzhi Li, Bei Yang, Heng Weng, Aihua Ou, Huixiao Hong, Zhaoxian Zhou, Ping Gong, and Chaoyang Zhang. 2017. Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC Bioinformatics* 18, 14 (2017), 523.

[250] Yu Wang, JunPeng Bao, JianQiang Du, and YongFeng Li. 2020. Precisely predicting acute kidney injury with convolutional neural network based on Electronic Health Record data. arXiv:2005.13171. Retrieved from https://arxiv.org/abs/2005.13171.

[251] Yu Wang, Peng-Fei Li, Yu Tian, Jing-Jing Ren, and Jing-Song Li. 2016. A shared decision-making system for diabetes medication choice utilizing Electronic Health Record data. *IEEE Journal of Biomedical and Health Informatics* 21, 5 (2016), 1280–1287.
[252] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 1322–1328.
[253] Hongyu Guo and Herna L. Viktor. 2004. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *ACM SIGKDD Explorations Newsletter* 6, 1 (2004), 30–39.
[254] Xiaoqian Jiang, Robert El-Kareh, and Lucila Ohno-Machado. 2011. Improving predictions in imbalanced data using pairwise expanded logistic regression. In *Proceedings of the AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association, 625.
[255] Hyun Kang. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology* 64, 5 (2013), 402.
[256] Peter Cummings. 2013. Missing data and multiple imputation. *JAMA Pediatrics* 167, 7 (2013), 656–661.
[257] Donald B. Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.
[258] Stef Van Buuren. 2018. *Flexible Imputation of Missing Data.* CRC press.
[259] Brian J. Wells, Amy S. Nowacki, Kevin Chagin, and Michael W. Kattan. 2013. Strategies for handling missing data in Electronic Health Record derived data. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)* 1, 3 (2013), 1035. DOI: http://dx.doi.org/10.13063/2327-9214.1035
[260] M. Smuk, J. R. Carpenter, and T. P. Morris. 2017. What impact do assumptions about missing data have on conclusions? A practical sensitivity analysis for a cancer survival registry. *BMC Medical Research Methodology* 17, 1 (2017), 21.
[261] MIT Critical Data. 2016. *Secondary Analysis of Electronic Health Records*. Springer Nature.
[262] Maria E. Montez-Rath, Wolfgang C. Winkelmayer, and Manisha Desai. 2014. Addressing missing data in clinical studies of kidney diseases. *Clinical Journal of the American Society of Nephrology* 9, 7 (2014), 1328–1335.
[263] Michael Lewis-Beck, Alan E. Bryman, and Tim Futing Liao. 2003. *The Sage Encyclopedia of Social Science Research Methods*. Sage Publications.
[264] Peng Li, Elizabeth A. Stuart, and David B. Allison. 2015. Multiple imputation: A flexible tool for handling missing data. *Jama* 314, 18 (2015), 1966–1967.
[265] Brett K. Beaulieu-Jones and Jason H. Moore. 2017. Missing data imputation in the Electronic Health Record using deeply learned autoencoders. In *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific, 207–218.
[266] Daniel J. Stekhoven and Peter Bühlmann. 2012. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.
[267] Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. 2017. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 787–792.
[268] Hadi Akbarzadeh Khorshidi, Uwe Aickelin, Gholamreza Haffari, and Behrooz Hassani-Mahmooei. 2019. Multi-objective semi-supervised clustering to identify health service patterns for injured patients. *Health Information Science and Systems* 7, 1 (2019), 18.
[269] Jukka K. Rönneikkö, Matti Mäkelä, Esa R. Jämsen, Heini Huhtala, Harriet Finne-Soveri, Anja Noro, and Jaakko N. Valvanne. 2017. Predictors for unplanned hospitalization of new home care clients. *Journal of the American Geriatrics Society* 65, 2 (2017), 407–414.
[270] Daniel J. Morgan, Bill Bame, Paul Zimand, Patrick Dooley, Kerri A. Thom, Anthony D. Harris, Soren Bentzen, Walt Ettinger, Stacy D. Garrett-Ray, J. Kathleen Tracy, and Yuanyuan Liang. 2019. Assessment of machine learning vs standard prediction rules for predicting hospital readmissions. *JAMA Network Open* 2, 3 (2019), e190348–e190348.
[271] Rolf H. H. Groenwold. 2020. Informative missingness in Electronic Health Record systems: The curse of knowing. *Diagnostic and Prognostic Research* 4, 1 (2020), 1–6.
[272] Katherine J. Lee, Gehan Roberts, Lex W. Doyle, Peter J. Anderson, and John B. Carlin. 2016. Multiple imputation for missing data in a longitudinal cohort study: A tutorial based on a detailed case study involving imputation of missing outcome data. *International Journal of Social Research Methodology* 19, 5 (2016), 575–591.
[273] Craig D. Newgard and Roger J. Lewis. 2015. Missing data: How to best account for what is not known. *Jama* 314, 9 (2015), 940–941.
[274] Sebastien Haneuse, Andy Bogart, Ina Jazic, Emily O. Westbrook, Denise Boudreau, Mary Kay Theis, Greg E. Simon, and David Arterburn. 2016. Learning about missing data mechanisms in Electronic Health Records-based research: A survey-based approach. *Epidemiology (Cambridge, Mass.)* 27, 1 (2016), 82.
[275] Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

[276] Glen B. Taksler, Jarrod E. Dalton, Adam T. Perzynski, Michael B. Rothberg, Alex Milinovich, Nikolas I. Krieger, Neal V. Dawson, Mary J. Roach, Michael D. Lewis, and Douglas Einstadter. 2021. Opportunities, pitfalls, and alternatives in adapting Electronic Health Records for health services research. *Medical Decision Making* 41, 2 (2021), 133–142.

[277] Nicolas Garcelon, Anita Burgun, Rémi Salomon, and Antoine Neuraz. 2020. Electronic Health Records for the diagnosis of rare diseases. *Kidney International* 97, 4 (2020), 676–686.

[278] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Proceedings of the Machine Learning for Healthcare Conference*. PMLR, 249–269.

[279] Neal A. Chatterjee, Jani T. Tikkanen, and Christine M. Albert. 2020. The electrocardiogram and sudden death: Capturing electrical physiology and arrhythmic substrate. *European Heart Journal* 41, 30 (2020), 2911–2912.

[280] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. 2018. Risk prediction on Electronic Health Records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1910–1919.

[281] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. 178, 11 (2018), 1544–1547. DOI: http://dx.doi.org/10.1001/jamainternmed.2018.3763

[282] Sherri Rose. 2018. Machine learning for prediction in Electronic Health data. *JAMA Network Open* 1, 4 (2018), e181404–e181404.

[283] Jens Christian Bjerring and Jacob Busch. 2020. Artificial intelligence and patient-centered decision-making. *Philosophyand Technology* 34 (2020), 1–23.

[284] Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics* 13 (02 2021), e1549. DOI: http://dx.doi.org/10.1002/wics.1549

Received November 2020; revised September 2021; accepted September 2021