Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 15-016

Mining Electronic Health Records (EHR): A Survey

Pranjul Yadav, Michael Steinbach, Vipin Kumar, Gyorgy Simon

October 13, 2015

Mining Electronic Health Records (EHR): A Survey

Pranjul Yadav, Michael Steinbach, Vipin Kumar, Gyorgy Simon

October 12, 2015

Abstract

The continuously increasing cost of the US healthcare system has received significant attention. Central to the ideas aimed at curbing this trend is the use of technology, in the form of the mandate to implement electronic health records (EHRs). EHRs consist of patient information such as demographics, medications, laboratory test results, diagnosis codes and procedures. Mining EHRs could lead to improvement in patient healthcare management as EHRs contain detailed information related to disease prognosis for large patient populations. In this manuscript, we provide a structured and comprehensive overview of data mining techniques for modeling EHR data. We first provide a detailed understanding of the major application areas to which EHR mining has been applied and then discuss the nature of EHR data and its accompanying challenges. Next, we describe major approaches used for EHR mining, the metrics associated with EHRs, and the various study designs. With this foundation, we then provide a systematic and methodological organization of existing data mining techniques used to model EHRs and discuss ideas for future research. We conclude with a case study of patients diagnosed with Type 2 diabetes mellitus (T2DM).

Contents

1 Introduction

The continuously increasing cost of the US healthcare system has received significant attention. Central to the ideas aimed at curbing this trend are a mandate from the Health Information Technology for Economic and Clinical Health (HITECH) Act to implement electronic health records (EHRs) and the ongoing transition from the current fee-for-service payment model to a new model based on population health management. Under this new model, primary care providers are responsible for managing entire patient populations with their payments tied to care quality. They are now incentivized to improve their efficiency by reducing wasteful treatments and laboratory tests.

One way to fulfill these goals is through unprecedentedly detailed disease prognosis and risk models. The enabling technology for this and other aspects of improved population health management lies in electronic health records which contain a wide range of patient information, including demographics, medications, laboratory test results, diagnosis codes and procedures, thereby capturing a more complete view of the medical state of a patient.

Thus, EHRs hold the promise of improving clinical quality while reducing healthcare costs. EHR data could serve as a foundation to create a learning health system: to rapidly develop new evidence,

translate the evidence into knowledge and apply the resultant knowledge to medical practice and health policy. EHRs have the potential to provide useful information to evaluate condition-specific clinical process metrics and outcomes, facilitate clinical decision support, enhance team-based population care outside the traditional face-to-face clinical encounter, and provide feedback on specific patient populations at the point of care. Multiple studies have observed that EHRs have reduced clinical errors [1–5], improved chronic illness care [6–10] and improved the completeness, accuracy and timeliness of case reporting to public health. EHRs provide unprecedented opportunities to identify genetic variants that influence susceptibility to common, complex diseases across geographies [11]. EHR-public health data exchange can provide superior public health surveillance information on chronic conditions such as asthma [12–15] and T2DM [10, 16–19]. It can also help in comparing risks to community factors such as economic disparity [20–22].

To better appreciate the opportunities associated with EHRs, we need to take a look at the traditional vehicle of clinical innovation, the Randomized Clinical Trials (RCTs) [23] which have hitherto been the gold standard for evaluating medical practices and treatments. RCT's are controlled studies that select patients for a study based on well-defined criteria and procedures such as random assignment of patients to the case and control groups. This randomization reduces bias in the patient subgroups by balancing known and unknown factors. Although RCTs have been in existence for many years, they suffer from some well-known limitations: without an EHR system from which to pull patient information they often have small sample size [24–30], confounding factors [31, 32], unaccounted for comorbidities[33], and short duration[34]. Although analyses based on EHR data will not replace randomized trials, they can help overcome some of these limitations and generate important health care insights. For example, with the help of EHRs, patient cohorts with minimal bias can be selected for RCTs and clinically significant findings from EHRs can be explicitly verified from RCTs. As a result, instead of being a substitute, EHRs are complementary to RCTs.

In the past, EHRs have been successfully used for various applications such as biomarker discovery, patient medical trajectories, measuring the efficacy of intervention, adverse event detection, pharmacovigilance, patient monitoring, sub-population analysis and measuring the effects of medical guidelines. For such applications, various techniques such as probabilistic graphical models, unsupervised clustering, classification approaches, and association rule mining have been widely used. In this manuscript, we provide a structured and comprehensive overview of the application of data mining techniques for EHRs as well as directions for future research.

Several articles, surveys and books have been written on EHRs [35]. Jensen et al. [36] discussed different types of EHR data including health records, radiological images and clinical texts. They presented how this data can be used for applications such as pharmacovigilance and subpopulation analysis. They also discussed limitations associated with EHRs such as patient privacy, patient consent and interoperability across institutions and countries. Murdoch et al. [37] discussed the application of big data to healthcare highlighting the various opportunities and limitations of using EHRs to improve the quality and efficiency of health care delivery. Hripscak et al. [38] presented various challenges associated with EHRs such as completeness, accuracy, complexity and bias. They also discussed how phenotyping can be used to overcome the aforementioned limitations. Correvita et al. [39] described various opportunities associated with EHRs such as the availability of larger samples, well defined variables and comparisons of results across institutions and geographical borders. They also discussed some limitations associated with EHRs such as the limited number of patient cohorts for important studies, complicated patient rights, consent management and semantic inter-operatbility. Fabricio [40] considered various challenges associated with EHRs in biomedicine, including the storage, transfer and security of patient information. Ross et al [41] analyzed EHRs in context of big data. They briefly

discussed certain application areas of EHRs such as pharmacovigilance, phenotyping and NLP. They also examined clinical decision support models, privacy and security associated with EHRs. Reddy and Agarwal [42] provided a comprehensive overview of different aspects of healthcare data analytics. In their book, they explored areas such as biomedical image analysis, sensor data analysis, biomedical signal analysis, genomic data analysis, clinical text mining and social media analysis. They also reviewed advanced topics such as clinical prediction models, visual analytics, clinico-genomic data integration, healthcare analytics for pervasive health, fraud detection and mobile imaging for biomedical applications.

Overview Following a description of application areas in Section 2, we describe EHR data in Section 3 and its associated challenges in Section 4. Given the unique nature of EHR data, these challenges have been addressed in data mining only to a very limited extent. In Section 5, we introduce techniques developed in other fields, most notably epidemiology and biostatistic, that address many of these challenges, although much more remains to be done. Section 6 introduced metrics that are used to measure the outcome of a treatment or study, while Section 7 describes study designs, which are the cornerstone of EHR data mining. A study design provides guidance on how the disparate collection of data tables comprising EHR data can be meaningfully organized, for example, how time can be handled, whether causation can be established and which metrics can be estimated. In Section 8, we turn our attention back to data mining and its contribution to the field of EHR mining. We provide a comprehensive overview of how data mining methods have been applied to mine EHR data. Finally, in Section 9, we discuss our findings and we conclude this survey with a case study that successfully answers complex clinical questions and serves as an illustration of how study design, data challenges and data mining methods interact.

2 Application Areas

While the fundamental question in medicine is to decide on the treatment that is most appropriate and effective for a particular patient, scientific inquiries into a disease typically start with the most basic epidemiological questions: How many patients at a particular time are affected by this condition? How many new cases do we discover each year? What are the symptoms of the disease? What is the *natural history of the disease*, i.e. what are the *precursors and consequences* of this condition?

Once we understand these fundamentals, we can try answering more advanced questions. To this end, we may need to assemble a cohort (group) of patients, some of whom are extremely likely to have the diseases (*cases*) and other who most likely do not (*controls*). This can be achieved through phenotyping algorithms, either hand-crafted or machine learned that characterize the disease in terms of patient characteristics observable from the EHR data. This problem is referred to as *cohort identification.*

For the cohort we could collect relevant known or potential predictors and build predictive models. These models predict the risk of disease, e.g. probability of developing the condition in 5 years (*risk prediction*) or investigate which predictors are relevant (*biomarker discovery*, *risk factor discovery*) in developing the outcome. Accurate knowledge of risk factors can help guide preventive efforts or focus interventions.

Interventions are often drug therapies or surgeries, but can also include recommendations for life style changes and/or patient education. Choosing the optimal treatment for a patient requires us to be able to estimate the effect of the possible interventions. Specialized data mining methods such as uplift modeling or statistical techniques in combination with causal analysis can be used to *quantify*

the effects of interventions.

Once the effect of a treatment has been proven in practice, this knowledge can be codified into and disseminated as clinical practice guidelines. *Evidence-based clinical practice guidelines* are considered a cornerstone of modern medicine, and they give guidance on the "optimal" treatment under a particular set of conditions based on epidemiological evidence. Guidelines are traditionally expert-crafted. The increasing role of computerized clinical decision support allows for more accurate but complex guidelines, suggesting that data mining technologies will play a more significant role in guideline construction.

While interventions typically help patients, occasionally they can lead to unforeseen events that adversely affect patient health such as surgical site infection or the unexpected reaction of multiple drugs. Predictive modeling has been used to both detect and to predict such adverse events.

In the following sections, we describe each of these applications in more details and discuss the role data mining has played thus far.

2.1 Understanding the Natural History of Disease

Analyzing and exploring disease statistics are often the first foray into a new disease or epidemiological study. These are aimed at answering questions such as the following: Is the condition or disease serious? Are large numbers of patients involved? What are the societal implications of the disease? Has the disease been studied before? These questions often prompt further investigation with rigorous study designs. Most of the time, the focus is on the prevalence of the disease, comorbidity analysis, or the incidence of the disease (patient medical trajectories).

Comorbidity analysis is the process of exploring and analyzing relationships between frequently co-occurring diseases. For example, patients suffering from type 2 diabetes mellitus (T2DM) often also suffer from hypertension, hyperlipidemia and impaired fasting glucose (IFG). Some diseases occur in clusters and it is desirable to treat them simultaneously. Further, analyzing the comorbidities and discovering the relationships among them, can lead to the modification of existing comorbidity scores (such as Charlson index) or to the development of novel ones. In the past, researchers have used comorbidity analysis for observing how alcohol usage is associated with depression, anxiety and personality disorders [43]. Doshi-Velez et al. used patient stratification techniques to observe comorbidities in patients suffering from autism spectrum disorders [44]. Wright et al. [45] employed the Apriori framework to detect associations between clinical concepts (laboratories test results, medications) and problem lists. Cao et al. [46] used a statistical framework to detect an association between diseases such as 'myasthenia gravis 'and 'cushingoid facies'. Similarly Holmes et al. [47] studied the comorbidities for rare diseases such as Kaposi sarcoma, toxoplasmosis, and Kawasaki disease. Using the association rule mining framework, Shin et al. [48] explored the comorbidities associated with hypertension such as non-insulin dependent diabetes mellitus, cerebral infection and chronic renal failure. Dasgupta et al. [49] analyzed disease drug relations by using advanced network clusters. They hypothesized that studying drugs in isolation can provide a different perspective on how two drugs can interact.

The medical state of a patient can be represented using laboratories test results, diagnosis codes or medication information, while the progression of a patient's medical state over time is known as the patient's *medical trajectory*. Examples of such trajectories are the progression of the patient from a healthy state through conditions like hypertension, hyperlipidemia, impaired fasting glucose (IFG), type 2 diabetes mellitus and eventually towards diabetes complications (e.g. amputation, severe paralysis or death). Often, multiple trajectories lead to the same outcome. For example, consider an

outcome such as mortality. In this case, a patient might die due to kidney complications, cardiovascular complications or peripheral complications. Even though the outcome is the same, the paths leading there are different. Research studies have observed that such varying trajectories can have significantly different associated risks for the same outcome. Examining such varying trajectories can lead to the development of tailored treatments, the discovery of biomarkers or the development of novel risk estimation indices.

Jensen et al. [50] have explored temporal disease progression patterns in data from an electronic health record registry which covers the entire population of Denmark. Using this cohort, they identified 1171 significant trajectories. These significant trajectories were then clustered using key diagnosis codes such as chronic obstructive pulmonary disease (COPD) and gout. Their findings demonstrate how these trajectories have predictive potential and might be the basis for predicting the next probable step in disease progression. Their findings also elaborate the association and causality of certain diseases. They further demonstrated how the population-wide disease trajectory approach uncovers diagnosis linkages which might conflict with research based on the past epidemiological studies. Teno et al. [51] examined differences in the pattern of functional decline among persons dying of cancer and other leading non cancer causes of death. They observed how patients with cancer experienced an increased rate of functional impairment beginning as late as 5 months prior to death. Murtagh et al. [52] analyzed how patients diagnosed with diseases have increased morbidity and an increased risk of death from cardiovascular disease. They also demonstrated how this exploration might lead to better patient management, thereby providing optimal care for patients in the terminal phase of their disease.

2.2 Cohort Identification

Cohort identification is the identification of patient groups satisfying the required criteria. This identification is performed using EHR attributes such as laboratories test results, vitals, medications and ICD-9 diagnosis codes. Traditionally, cohort identification was carried out through chart reviews. However the scale enabled by EHRs render manual chart review impractical. Instead, electronic phenotyping algorithms are applied with manual spot-checking. Cohort identification has been widely used in various clinical research studies and biomedical applications. This process is often the platform for carrying out future studies in areas such as pharmacovigilance, predicting complications, and quantifying the effect of interventions.

Cohort identification usually employs supervised learning techniques, where the gold standard is defined using expert clinical knowledge. Such identification, using ICD-9 codes and narrative data, has been used to develop automated models to identify patients with cancer [53], rheumatoid arthritis [54], pneumonia [54], critical care [55] and asthma [56]. Kandula et al. [57] developed a bootstrapping learning method that, starting with an initial classification based on ICD-9 codes, iteratively improves cohort accuracy through training on relevant structured data. Their proposed method does not require prior information about the true class of the patients. They used their method to identify T2DM and hyperlipidemia patient cohorts from a database of 800,000 patients. Rasmussen et al. [58] discussed phenotype design patterns based on existing phenotype algorithm definitions from the eMERGE network. They believed it would help researchers in working with EHR data for algorithm development.

A phenotype is defined as a biochemical or physical trait of an organism, such as a disease, physical characteristic, or blood type, based on genetic information and environmental influences. Examples of phenotypes in EHRs are clinical conditions, characteristics or sets of clinical features that can be determined solely from the EHR data and do not require a chart review or interpretation by a clinician.

Such techniques are useful for identifying patients or populations with a given characteristic or condition of interest from EHRs using data that are routinely collected in EHRs or ancillary data sources such as disease registries or claims data. Phenotyping queries used for cohort identification can be replicated at multiple sites in a consistent fashion in order to ensure that populations identified from different healthcare organizations have similar features. Phenotypic definitions can also be used for direct identification of cohorts based on population characteristics, risk factors, and complications, allowing decision-makers to identify and target patients for screening tests and interventions that have been demonstrated to be effective in similar populations.

Castelli et al. [59] used phenotyping to analyze the relationship between coronary heart disease (CHD) prevalence and fasting lipid levels. They observed how inverse HDL cholesterol-CHD association was not appreciably diminished when adjusted for levels of low density lipoprotein (LDL) cholesterol and triglyceride. Newton et al. [60] worked on validating EMR-derived phenotypes and made the following observations: multisite validation improves phenotype algorithm accuracy, algorithm development and validation work best as an iterative process, validation by content experts or structured chart review can provide accurate results and patient movement in and out of the health plan (transience) can result in incomplete or fragmented data. Overby et al. [61] worked on developing a collaborative approach for an electronic health record (EHR) phenotyping algorithm for drug-induced liver injury (DILI) and demonstrated the portability of their algorithm across multiple institutions. They also observed that the performance of their algorithm for identifying DILI was comparable with other computerized approaches used to identify adverse drug events.

Pathak et al. [62] identified various challenge associated with phenotyping EHRs including developing approaches for high-throughput extraction and representation of phenotypes, building techniques for storing, integrating, and querying phenotype data and advancing phenotypic-driven analysis to derive phenotype-genotype associations. Schram et al. [63] worked on an extensive phenotyping study that focuses on the etiology of type 2 diabetes (T2DM), its associated complications, and its emerging comorbidities. Their study uses state-of-the-art imaging techniques and extensive biobanking to determine health status in a population-based cohort of several thousand individuals that is enriched with T2DM individuals. Boland et al. [64] introduced a new concept called verotype by integrating genetics along with EHRs for patient identification. They believed verotypes would be useful for personalized medical treatment regiments. Gotz et al. [65] combined data mining and visualization techniques to retrieve patient cohorts that satisfy complex clinical events. They achieved this by integrating visual queries, on-demand analytics and interactive visualization. Their system also provided an interactive visual environment for the exploration and analysis of temporal medical event data.

Wang et al. [66] worked on segmenting patient cohorts by incorporating prior knowledge from domain experts. They hypothesized that such domain knowledge is very important as it reflects crucial medical insights which are validated by extensive clinical studies. They then used these cohorts for developing group-specific risk prediction models. Peissig et al [67] developed a technique to identify subjects with age-related cataracts and the associated cataract attributes using only information amiable in the EHR. They demonstrated that a multi-modal approach which includes the use of EHRs along with clinical notes increases the predictive performance. Pathak et al. [68] proposed semantic web technologies for extracting phenotyping data from EHRs. They discussed how such techniques would allow federated querying, reasoning, and efficient information retrieval across multiple sources of clinical data and information. More recently, Ho et al. [69–71] proposed tensor factorization methods to derive phenotypes. Schulam et al. [72] proposed the Probabiliistic Subtyping Model (PSM) to identify subgroups based on clustering individual clinical severity markers. Their method uses hierarchical

clustering to account for variability arising due to noise and irregular sampling methods. Hu et al. [73] proposed a vector space model to represent patient utilization profiles, and apply clustering techniques to identify utilization groups within a given population. Their technique can be used to identify high utilization users from low utilization users thereby leading to detection of anomalous patient profiles.

2.3 Risk Prediction/Biomarker Discovery

Risk prediction is the problem of constructing predictive models to assess the patient's risk and progression from a patient's current medical state to a medical state associated with potentially advanced medical complications. Such analysis is often performed to identify high risk individuals, thereby facilitating the design and planning of one's treatment plan [74–76]. Such analysis might lead to improvement in a patient's health, thereby preventing the patient from progressing to advanced complications. In some cases, predicting the patient's risk of progression is secondary to understanding the underlying risk factors. Risk models can provide information about the importance of risk factors.

With the availability of EHRs, models can be developed for assessing the patient's risk for multiple diseases. Such models also have the capability to capture effects arising due to demographic attributes such as age, gender, race, ethnicity and social status. Further due to the interoperability associated with EHRs, models can also be developed using data across geographies thereby incorporating the genetic makeup of the patients.

Data mining techniques such as logistic regression, Poisson regression and survival modeling techniques such as Cox proportional hazards regression are often used to analyze the patient risks for a complication of interest. Greenland et al. [77] analyzed how risk assessment associated with coronary heart disease might be improved by additional tests such as coronary artery calcium scoring (CACS). Knaus et al. [78] refined the APACHE (Acute Physiology, Age, Chronic Health Evaluation) methodology in order to more accurately predict hospital mortality risk for critically ill patients in ICUs. They also analyzed the relationship between the patient's likelihood of surviving to hospital discharge and the following variables: major medical and surgical disease categories, acute physiologic abnormalities, preexisting functional limitations, major comorbidities, and treatment location immediately prior to ICU admission.

Sarkar et al. [79] presented a methodology for developing an improved feature selection technique that will help in accurate prediction of outcomes after hematopoietic stem cell transplantation (HSCT) for patients with acute myelogenous leukaemia (AML). They also observed how their selected features were similar to those obtained by traditional statistical techniques. Letham et al. [80] used Bayesian model and Markov chain Monte Carlo sampling to develop interpretable predictive models using EHRs data. Ebadollahi et al [81] worked on developing a decision support tool for near-term prognostic insight to help clinicians better assess the impact of their decisions. They used inter-patient similarity to project patient data into the future to provide insights about the query patient. Feldman and Chawla [82] presented ADMIT (Admission Duration Model for Infant Treatment) model, which yields personalized length of stay estimates for an infant, utilizing data available from time of admission to the ICU. Their algorithm utilizes an augmentation of the Adaptive Boost algorithm, known as the LogitBoost. Ngufor et al. [83] developed an efficient and accurate algorithm that could estimate the risk of multiple outcomes simultaneously such as perioperative bleeding, intraoperative RBC transfusion, ICU care, and ICU length of stay. Byrd et al. [84] constructed a system to automatically identify heart failure diagnostic criteria. Kamkar et al. [85] used Tree Lasso for feature selection along with state of the art classification problems for identifying stable risk factors for many healthcare problems. Tran et al. [86] worked on development of auto-extracted standard features from complex medical records, in

a disease and task agnostic measure. They demonstrated how their auto-extracted features achieve better discriminative power for prediction hospital readmission.

Lakshmanan et al. [87] an approach for mining clinical care pathways correlated with patient outcomes that involves a combination of clustering, process mining and frequent pattern mining. Wang et al. [88] proposed a probabilistic disease progression model that continuously learns from discrete-time observations with non-equal intervals. Their model is also capable of learning full progression trajectory. They demonstrated the applicability of their model on diseases such as T2DM, cardiovascular and psychological complications. Vijayakrishnan et al. [89] analyzed EHRs for earlier identification of disease states such as heart failure (HF). They developed a novel text and data analytic tool for analyzing longitudinal EHRs of over 50,0000 primary care patients. Vellanki et al. [90] used Bayesian Nonparametric factor analysis along with clustering techniques to identity biomarkers for children diagnosed with autism spectrum disorder (ASD). They demonstrated that by using bayesian nonparametric framework, once can discoed learning patterns more efficiently as compared to the parametric methods.

Risk prediction also provides the opportunity to identify significant indicators of a biological state or condition. In simple terms, a biomarker is defined as a set of measurable quantities that can serve as an indicator of a patient's health. For example, abnormal hemoglobin A1C is a biomarker for T2DM and hyperlipidemia is a biomarker for being at risk of cardio-vascular complications. Similarly, there are certain biomarkers, which are common across many diseases. For example, age is by far the most common biomarker. It indicates that as a person ages, his or her risk to acquire certain diseases (e.g T2DM, cardio-vascular complications and kidney complications) increases. EHRs provide a platform to identify, analyze and explore biomarkers for different diseases.

Biomarkers offer a succinct summary of the patient's state with respect to a medical condition. Rather than having to analyze the thousands of variables present in an EHR, it can be sufficient to focus on relatively few biomarkers to paint a reasonably accurate picture of the patient's overall health. Over the years, biomarkers have found numerous applications. They can be used in rule-based systems to identify cohorts for a clinical trial or to enhance existing risk indices (e.g. Framingham risk scores). Data mining techniques have been extensively used to discover biomarkers. Schrom et al. used association rule mining along with propensity score matching to investigate how statin can lead to overt diabetes in certain subpopulations [91]. Supervised techniques, such as survival association rule mining [92], have been used to discover biomarkers for T2DM. An example found by the technique is the combination of (hyperlipidemia, triglycerides and fibrates), which indicates high relative risk for T2DM. Harpaz et al. [93] used statistical techniques to identify a chemical biomarker rasburicase, that results in adverse events related to pancreatitis.

2.4 Predicting the next complication: What and When

Future complications arising due to patient's current medical condition can be classified into two categories, i.e., short term and long term complications. This becomes more relevant as patients follow multiple health trajectories, often leading to life threatening conditions including mortality [94, 95]. For clinical purposes, predicting the next complication is a challenging problem. For example, predicting the onset of neonatal sepsis, rehospitalization or the next complication for a patient diagnosed with T2DM.

The availability of large patient cohorts across longer observation periods provide an opportunity to build clinical decision support models that can be used to predict complications across multiple

observation periods. Such models would be quite useful for assessing the risk caused by diseases such as T2DM, where it takes around 5-10 years for a patient to progress from one state of complication to another potentially advanced complication. Clinical decision support systems informed by such models will lead to improved patient care, thereby improving overall health care. Such analysis can lead to the development of personalized and tailored interventions. It also can lead to refinement of medical guidelines by considering the short and long term impact.

Generalized linear regression and survival modeling techniques such as Cox proportional hazards regression are often used for the development of such models. Yadav et al. [96] used Cox proportional hazards regression to estimate the risk of potentially advanced complications such as Peripheral Vascular Disease (PVD), Cerebral Vascular Disease (CVD), Ischemic Heart Disease (IHD) and Congestive Heart Failure (CHF), often associated with T2DM. They first developed a diabetes complication index which summarizes a patient's health in terms of post-diabetic complications into a single score. Through the use of this score, they track a patient's health and show that distinct trajectories in diabetes can be identified thereby demonstrating the need and laying the foundation for future clinical EBP guidelines that take trajectories into account. Zhao et al. [97] proposed a novel method which combines PubMed knowledge and EHRs to develop a weighted Bayesian Network Inference (BNI) model for pancreatic cancer prediction. Their model was further used to compute probabilities for various risk factors or complications associated with pancreatic cancer prediction. Considerable research has been performed to predict future complications associated with lung cancer [98], cardiac arrest [99], bariatric surgery [100], carotid endarterectomy [101], acute cough [102], breast reconstruction [103], pulmonary resection [104–107], knee replacement [108], lumbar decompression [109], orthopedic surgery [110], hysteroscopic surgery [111] and febrile neutropenic cancer [112].

Lui and Hauskrecht [113–116] modeled the irregularly sampled clinical time series by using multiple Gaussian process sequences in the lower level of our hierarchical framework and capture the transitions between Gaussian processes by utilizing the linear dynamical system. They used their technique for exploring complications associated with complete blood count (CBC) panel data for post-surgical cardiac patients during hospitalization. Panahiazar et al. [117] used EHRs for inferring an individual patient's response to Heart Failure therapy. To carry out the aforementioned objective, they used patient-specific information from the EHR, including medical comorbidities, laboratory measurements, ejection fraction, vital status and demographics to identify similar patients.

2.5 Quantifying the effect of Intervention

Life-style modifications such as smoking cessation, low-calorie food consumption and medication prescriptions are common examples of medical interventions. Such interventions are usually advised when the patient has a high probability of progressing to high-risk complications (e.g. mortality). The ability to quantify the effect of interventions can lead to development of sophisticated and tailored medical treatments.

The longitudinal aspect of EHRs provides an opportunity to analyze the effects of intervention for longer period of time across larger cohorts. It also provides clinicians with a platform to analyze whether the interventions have any accompanying adverse effects. Moreover, EHRs provide a platform to analyze whether interventions vary across cohorts based on gender, age, ethnic make-up, socio-economic status, etc.

Data mining techniques such as association rule mining have been used to measure the effect of interventions. Statin is an example of a commonly prescribed medication for patients diagnosed with

hypercholesterolemia. Schrom et al. [91] used association rule mining along with propensity score matching to identify how the use of statin leads to overt diabetes in certain sub-populations. The sub-population consisted of patients diagnosed with hypercholesterolemia. They demonstrated their technique on a real diabetes data set by examining the relationship between statin use and diabetes, and identified novel risk factors. Campbell [118] analyzed a variety of tests and measures that are useful in documenting and quantifying the outcomes of intervention for persons with cerebral palsy. Proschaska et al. [119] analyzed the effects of risky behaviors such as smoking, alcohol abuse, physical inactivity, and poor diet on human health. Ronsmans and Campbell [120] presented the evidence of the effect of health interventions on mortality reduction from hypertensive diseases in pregnancy. Law et al. [121] analyzed how statins reduce serum concentrations of low density lipoprotein (LDL) cholesterol and the incidence of ischemic heart disease (IHD) events and stroke, according to drug, dose, and duration of treatment.

2.6 Constructing Evidence Based Guidelines

Clinical guidelines are systematically developed descriptive tools or standardized specifications for care to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances [122]. Evidence based guidelines (EBG) try to guide decision making by identifying best clinical practices, that are meant to improve the quality of patient care [123]. They help clinicians make sound decisions by presenting up to date information about best practices for treating patients in a particular medical state including expected outcomes and recommended follow up interval. For example, EBG guidelines for diabetes consist of rules such as symptom identification checks (e.g. diagnosis of T2DM when fasting plasma glucose is greater than 7), lifestyle modification recommendations (e.g. cessation of smoking), medication order (e.g. prescription of metformin), etc. These guidelines are often regarded as the cornerstone of modern healthcare management.

Designing effective EBGs requires a large enough sample of the target population with long follow-up duration to study the outcome. Since EHRs often satisfies these criteria, they can be very useful in evaluating medical guidelines. Data mining techniques such as association rule mining, sequential rule mining and regression approaches can be used to develop and test existing guidelines[124],[125]. EBG's have been developed for diseases treated in the emergency department [126], medication therapy for upper respiratory tract infection [127], ear, nose, diabetes mellitus type 2 (T2DM) [128], prosthodontics [129], etc. Guidelines might vary across geographies due to differences in population genetics, life-style and socio-economic status. For example, different sets of guidelines have been developed for T2DM by Finland [130] and Singapore[131]). Pivovarov et al. [132] analyzed the potential overuse of certain clinical guidelines. In particular they looked at hemoglobin A1c testing across 119 000 patients and 15 years of hospital records. They also examined the patterns before A1c was included in American Diabetes Association guidelines. Their study demonstrated over utilization of A1c and attributed this to lack of care coordination and point of care tests followed by confirmatory laboratory tests.

2.7 Adverse Event Detection

Adverse event detection refers to the problem of detecting any untoward medical occurrence caused by mismanagement of patient health. Such medical errors might arise due to accidental surgical practices, drug reactions or the use of outdated medical guidelines. Examples of such events are detecting patients with high risk for narcotic dosing error, assessing the disagreement between the medication order and medication delivery and identifying fatal events in ICU [133]. Identification of such events are not only important to the patient (medical health), but also to the healthcare provider (in terms of cost reduction). Moreover, analysis of such events might lead to the review of antiquated guidelines,

withdrawal of certain drugs (those causing adverse events) from the market, etc. We categorize the research associated with adverse events into two major areas i.e. pharmacovigilance and patient monitoring.

Pharmacovigilance deals with monitoring and detecting the adverse effects of medications/drugs. Such adverse drug reactions can prove fatal to patients and can also have a significant impact on healthcare management. Although drugs are tested for any potential adverse effects before they are released for widespread use, often test cohorts are small with short observation periods. Several agencies conduct research on detecting ADRs: FDA with its adverse event reporting system, the European Medicines Agency, and the World Health Organization, which maintains an international adverse reaction database. Despite this effort, all these agencies suffer from underreporting and biased analyses of adverse drug reactions. EHRs provide a new platform to improve and complement drug safety surveillance strategies. Extensive research on ADRs has been performed in the context of cardiovascular complications, pancreatic complications [93, 134] and allergies [135]. Supervised techniques, such as disproportionality analysis, logistic regression, Bayesian inference and NLP techniques have frequently been used to discover ADRs [136, 137]. Besides EHR data, some research has also used weblogs to identify ADRs [138]. Similarly, research has been conducted using statistical analysis to identify certain medications which lead to adverse effects [137–139]. Bobo et al. [140] designed an algorithm to identify new or prevalent users of antidepressant medications via population-based drugs-pescription records and confirmed that prescription records can be used to identify prevalent or incident users of antidepressants. Pathak et al. [141, 142] used Semantic Web and Linked Data technologies for identifying potential drug-drug interaction (DDI) information from publicly available resources, and determining if such interactions were observed using real patient data. Specificially, they analyzed widely prescribed cardiovascular drugs: Warfarin, Clopidogrel and Simvastatin. Sathyanarayana et al. [143] proposed a data driven framework to predict the effectiveness of medication on a patient diagnosed with T2DM. Their aim was to evaluate the effectiveness of Metformin. Decisions trees and random forests were used for their analysis.

Patient Monitoring Surveillance is the continuous monitoring of patients by using diverse information such as biochemical markers (e.g. glucose, hemoglobin A1C, and blood urea nitrogen), voice analysis, physiological variables (e.g. heart rate, breathing rate, heart rate variability, and sleeping alterations) and behavioral data (e.g. stress related hormones and activity recognition). Round the clock monitoring helps clinicians explore and understand the causal factors responsible for adverse events. Such surveillance helps analyze large patient cohorts with limited clinical support, patient health management during critical times (depressive and maniac episodes), etc. Surveillance techniques are frequently used for patients admitted to the ICU. Nachimuthu et al. [144] used them to identify fluctuations in glucose levels while Rose et al. [145] used them for patients admitted for hemodialysis. Such techniques can lead to a reduction in overall health care costs, access barriers [146, 147], unnecessary hospital admissions, frequency of primary care visits and improvement in illness prevention and care co-ordination [148].

3 Nature of EHR Data

One motivation behind the federal mandate for EHRs was to document patients' state of health over time and the therapeutic interventions to which these patients were subjected. EHRs store this information in structured (databases), semi-structured (flow sheets) and unstructured formats (clinical notes). The format of the information greatly affects the ease of access and quality of the data, and thus has substantial impact on the downstream data mining.

3.1 Structured Data

From the viewpoint of healthcare analytics, retrieving structured data is the most straightforward. Structured data is stored in database tables with a fixed schema designed by the EHR vendor. The most commonly used information, such as demographic information (e.g. birth date, race, ethnicity), encounters (e.g. admission and discharge data), diagnosis codes (historic and current), procedure codes, laboratory results, medications, allergies, social information (e.g. tobacco usage) and some vital signs (blood pressure, pulse, weight, height) are all stored in structured tables. This kind of information is common across providers and not specific to any clinical specialty. Thus the use and format of this information is well handled by the EHR vendors. This allows such information to be stored in structured data tables with apriori defined layouts (schema). Fixed schemas enable high performance (rapid access to data) and standardization: the schemas for these tables are very similar if not identical across installations by the same EHR vendor, requiring very little (if any) site-specific knowledge from users. This quasi-standardization of fields also greatly helps information retrieval for analytic purposes.

Storing all information in EHRs as structured elements, however, is impractical: it would require anticipation of all possible data elements (e.g. metrics whose usefulness we do not yet appreciate) and would result in a level of complexity that would render the EHR system unusable. However, there is a need for storing information that does not readily fit into the admittedly rigid schema of the structured tables. For example, clinicians often write notes about patient's symptoms based on their previous experiences, which is hard to standardize apriori.

3.2 Unstructured Data

Among the three formats, clinical notes (unstructured data) offer maximal flexibility. Clinical notes mostly store narrative data (free text). Many types of clinical notes are in existence, and the type of note (e.g. radiology report, surgical note, discharge notes) is the only limiting factor on the type and breadth of information the note in question can store. Information regarding a patient's medical history (diseases as well as interventions), familial history of diseases, environmental exposures and lifestyle data all reside in clinical notes. Natural language processing (NLP) tools and techniques have been widely used to extract knowledge from EHR data.

Clinical notes such as admission, treatment and discharge summaries store valuable medical information about the patient, but these clinical notes are very subjective to the doctor or the nurse writing them, and lack a common structure or framework. These clinical notes an also have grammatical errors, short phrases, abbreviations, local dialects and misspelled words. Considerable data processing needs to be conducted on these clinical notes such as spelling correction, word sense disambiguation, contextual feature detection, extraction of ICD codes from clinical text, and adverse events surveillance. This makes deriving structured information about patient phenotypes from clinical notes a computationally challenging task that requires the most sophisticated NLP tools and techniques. In the past, work has been done to analyze the effect of time constraints on routine clinical tasks such as review of ambulatory EHR clinical notes [149], creation of clinical sense inventory of clinical abbreviations and acronyms [150] and development of tailored NLP methods to extract information from operative notes [151].

3.3 Flowsheets

In between the two extremes (structured tables and unstructured clinical notes) lies the (semi-structured) flow sheet format [152]. This format is most reminiscent of resource description files (RDF), consisting of name, value and time stamp triplets. Typically, the "name" field stores the name of the measure and the "value" field contains the actual measurements: e.g. the name is "arterial blood pressure" and the value is 145 Hgmm. This format is more flexible than the structured tables, since the user can define new metric through the name field; the set of metrics is not restricted to those anticipated by the EHR vendor. Flow sheets are similar to structured data in the sense that the value field is either a quantitative measure (e.g. blood pressure) or typically a restricted set of values. For instance, the American Society of Anesthesiologists (ASA) physical status takes values of "healthy", "mild systemic disease", "severe systemic disease", "severe life-threatening systemic disease", or "moribund".

Flow sheets offer expandability to EHR systems and thus have found numerous uses, becoming the only or most convenient data repository for many applications. Possibly the most important use for flow sheets is that they provide detailed information about specialty care. For example, information related to a patient's asthma care plans can be stored in flow sheets or they may store various diabetes-related non-standard (or not-yet-standard) metrics for a diabetes clinic. In addition, they may provide additional details regarding how a particular measure was obtained (blood pressure taken while the patient was lying flat) and can also be used to store automated sensor data (e.g. pulse and blood oxygen levels every few minutes in an intensive care unit). Further, flow sheets can be used to pull together related measurements such as quality indicators.

4 Data-Related Challenges

EHR data as a research platform poses numerous challenges including data integration across multiple provider sites each with its own best practice and across multiple sources such as clinical, claims and high-dimensional data. In the remainder of this section, we examine a number of data related challenges, all of which are, at least in part, a consequence of missing data. We provide an overview of those challenges here, and more details in the rest of this section.

The single most important challenge, the one that arguably impacts data mining methodologies at the most, is missing data. EHR data can be missing for a wide variety of reasons. First, as we have discussed earlier, it is study designs that transform raw EHR data into a design matrix, a matrix that is amenable to the application of data mining techniques. Each study design defines a study period, a time period during which the patient is under observation. Events that take place outside the study period are unobservable and data that is unobservable because it falls outside the study period is referred to as censoring.

Even during the study period, patients' health state is not always observable. The US health-care system allows patients to seek medical care from multiple providers who are not required to exchange health information. Fragmentation refers to the situation when a patient's trajectory is only partially observable during the study period, because the patient sought care at a different provider who did not share data with those conducting the study. Naturally, all diagnoses, tests and treatments received at the other provider are unobservable. This is known as fragmentation.

Missing data can still arise when the patient receives unfragmented care from a single provider, simply because the patient receives care intermittently. During every patient care encounter, providers

focus on a limited set of ailments, and hence update only a small fraction of the patient's record. The irregular nature of the visits and the small fraction of the record that gets updated during any visit leave a large portion of the record unobserved for extended periods of time unobserved or in other words, missing. We refer to this as irregular data.

Even during a single encounter, not all information about the patient gets recorded. EHRs are notoriously lacking in terms of documenting socio-economic data, environmental exposures and lifestyle descriptors. Unobserved descriptors and lack of knowledge about disease processes can lead to biases and confounding. It is not only "soft" socio-economomic and lifetime data that could be missing; "hard" medical facts, such as diagnosis codes, could also be missing. Indeed, the recording of diagnosis codes is often dictated by reimbursement rules.

In the following subsections, we will discuss these various issues in details.

4.1 Censored Data

By censored data, we refer to data for which information about a patient's medical state is observed only during a certain period of time or conversely, when potentially interesting events fall outside the observation period and are hence unobservable. In case of left censored data, patients experienced events of interest prior to the start of the study; in case of right censoring, potentially interesting events are unobservable because they happened to the patient after the study concluded. In case of interval censored data, information is only available of the data being within a certain limit. Studies can be either left, right or interval censored. Censoring can lead to loss of crucial information about the patient's health. For example, for the right censored patient, there is neither an easy way to determine whether the patient is alive or dead nor to measure the efficacy of the treatment the patient was undergoing. Examples of datasets with censoring are T2DM [153], dementia [154], nepropathy [155] and mortality [155].

Survival modeling techniques analyze data where the outcome variable is the time until the occurrence of an event of interest are frequently used to model censored data. Example of such techniques are nonparametric estimation methods such as Kaplan Meier curve [156], bayesian non-parametric methods [156] which involves prior belief about the shape of the survival function, semi parametric proportional hazards regression with fixed covariates or time dependent covariates [157], additive hazards regression model and parametric regression models using weibull distribution or log logistic distribution [158].

4.2 Fragmentation

Fragmentation is a lack of data sharing across providers. Fragmentation typically occurs when patients visit multiple healthcare providers seeking specialty care, expert advice or second opinions. In such scenarios, all healthcare provider involved only have partial information about the patient's medical history. Integrating data across multiple healthcare providers has several limitations. These challenges arise as different EHR systems such as General Electic (GE) or Epic, require a common language to transfer information into HL-7 (common protocol), which cannot capture all nuances. Even when multiple sites use the same EHR, their treatment policies may differ, flowsheets may differ and thus their definitions of nuanced concepts may differ. For example, fasting and random glucose measurements are not distinguished by lab codes and different sites can apply different methods to distinguish the two.

4.3 Irregular Time Series Data

Beside our inability to make observations before the study period starts or after it concludes, the most striking characteristic of the EHRs data is the irregularity of the patient visits. While recommended frequency of visits may exist, few patients actually follow these recommendations. For example, as per the ADA guideline A1C test must be performed at least two times a year for individuals who are meeting treatment targets and have stable glycemic control. On the contrary, an A1C test must be performed quarterly for individuals whose therapy has changed or who are not meeting glycemic targets.

Further, Information such as vitals is collected at every visit, certain laboratories tests are ordered annually, and other tests are performed only as needed. For example, as per the ADA guidelines, a laboratory test to measure Hemoglobin A1C in blood is recommended every six to twelve months, but a bacterial panel is ordered only when needed. This difference in the frequency of collection of medical information leads to irregular longitudinal data. This difference in information being captured from roughly every year to every few hours, leads to the problem of multiple temporal scales. Care must be taken to compare the trajectories of different health indicators, as they might have varying temporal scales. This irregular time gap between visits can be further complicated as patients often have different diseases with accompaning complications.

Analyzing regular time series is a well-studied problem in data mining, but application of these techniques to EHR-type irregular time series is very challenging.

4.4 Other Sources of Missing Data

Diagnosis codes might also be missing due to intentional omission as diagnosis codes, especially billing codes, are related to reimbursement. Different problems, comorbidities or complications have different reimbursement rates: depending upon the complications, the same procedure may have different costs and thus result in increased or decreased reimbursements. Due to these financial constraints, only some of the problems related to the primary cause of the visit, are used to generate billing codes (ICD codes are used to represent these problems in billing records). This leads to biases in ICD-9 codes, as the billing codes might not be a true representation of the actual medical state of the patient.

Diagnosis codes can also be missing due to changes in disease definitions and updates to the ICD-9 codes. For example, pre-diabetes did not have a corresponding ICD-9 code until 2000. The introduction of new and the periodic updates to existing ICD codes leads to further complications such as lack of a clear mapping from the old revision to the new and subsequently, to inconsistent research findings.

Another largely unobservable source of missing information lies in patient conformance with prescriptions and other physician advice, such as lifestyle change recommendations. The orders table in an EHR indicates that the physician prescribed a medication, but in most cases we do not know whether the patient actually took the medication. This situation is referred to as Intent to Treat. In the case of the lifestyle change, we may not even have documentation that the patient received this advice.

A unique aspect of missing data in clinical analytics is that whether the data is missing or not can be predictive. When a physician orders a test, he usually suspects that the patient may suffer from the corresponding condition. Conversely, by not ordering certain tests, the clinician suggests that corresponding medical conditions are absent. For example, no bacterial panel being ordered likely indicates that the patient is not suffering from any infection.

4.5 Consequences of Missing Data

Treating patients who need acute medical attention without a complete medical history can result in medical errors and redundant tests being performed. Incomplete patient history can delay the patient's treatment and inflate the cost of the treatment.

4.6 Biases and Confounding Effects

Studies performed using EHRs often have biases and confounding effects [159, 160]. Biases might arise due to multiple reasons. For example, in a cohort study, there might be significant differences in baseline characteristics (age, gender, race, ethnicity) between the cases and the controls [161–164]. In such cases, any observed difference between the groups after a follow-up period might be due to the difference in baseline characteristics and not due to the exposure. Therefore in such cases, analyzing the real effect of exposure might be difficult.

Such bias can be overcome by finding the right control group. One possible way is to randomly select subjects from a pool of patients such that the pool does not comprise of patients diagnosed with the outcome [165]. In other approaches, controls can be drawn from neighborhood of the cases as such controls would be very similar in terms of socio-economic status and lifestyle choices [166, 167]. Similarly, when genetic factors are the main focus of study, controls could often be chosen from family and relatives as they share similar genetic make-up [165].

Confounding is another issue which might undermine the internal validity of any study [168, 169]. Such situation arises when a variable (i.e. confounder) is associated with the exposure and affects the outcome, but the confounder variable is not an intermediate link in the chain of causation between exposure and outcome [170–172]. For example, studies have often reported a high degree of association between risk of myocardial infarction and oral contraceptives. However, it was later observed that this association was spurious because of the high proportion of tobacco users among users of birth control pills. Therefore tobacco consumption confounded the relation between oral contraceptives and myocardial infarction.

Multiple ways to overcome confounding effects have been proposed. The simplest strategy is to restrict or exclude subjects which might lead to confounding effects [169]. For example, if there are few subjects who consume tobacco, then it would help to remove these subjects from the study. Similarly, pairwise matching [169] and stratification [173] are also techniques used to avoid confounders. However, the techniques used until now are mostly used to avoid confounding arise due to single variable effects. To handle multiple confounding variables, multivariate modeling techniques can be used [174]. For example, survival modeling techniques such as Cox proportional hazards regression can be used to model time to death. Such methods might control simultaneously for age, blood pressure, smoking history and other risk factors.

The aforementioned challenges described in this section are unique to EHRs. We believe that substantial literature exists and the readers might be aware of the commonly associated data challenges such as noise, high dimensionality, sparseness, non-linear relationships, dependencies between various variables and wide variation in the data types. For detailed description related to missing data, please refer to our technical report yadav et al.

5 Approaches

In the previous section, we outlined a number of challenges that the EHR data poses for analysis. Addressing some of these, e.g. fragmentation, is a policy question and is not under the control of the investigator or the analyst. However, others, most notably censoring, the irregular temporal nature of the data, and bias and confounding, can be and have been addressed through various approaches. In this section, we provide an overview of these approaches.

5.1 Handling Censored Data

Censoring occurs when a patient's trajectory is only partially observable. For example, suppose a study is conducted to measure the impact of a diabetes related drug on mortality rate. In such a study, let us assume that the individual withdrew from the study after following the study course for limited duration. In such a scenario, information about patient's vital statistics is only available until the patient was censored. Such data is common in domains such as healthcare and actuarial work.

Survival analysis is an area of statistics that deals with censored data, such as death in biological organisms or failure in mechanical systems. These approaches usually aim to answer questions such as the following: what is the proportion of a population that will survive past a certain time? How do particular circumstances or characteristics increase or decrease the probability of survival? These techniques can be divided into three major categories: non-parametric, semi-parametric and parametric.

Non-parametric techniques do not rely on assumptions about the shape or parameters of the distribution of time to event. Examples of such techniques include Kaplan-Meier estimators [156] and Nelson- Aalen estimators [175]. Rihal et al. [176] used Kaplan-Meir estimators for incidence and prognostic implications of acute renal failure in patients undergoing percutaneous coronary intervention (PCI). Dormandy et al.[177] used Kaplan-Meier estimates in their analysis of patients who were diagnosed with T2DM and were at high risk of data and non-fatal myocardial infarction and stroke. Rossing et al. [178] used Nelson-Aalen estimators for analyzing the predictors of mortality in insulin dependent diabetes. et al. Similarly Ekinci et al. [179] used it for exploring salt intake and mortality in patients with type 2 diabetes.

Parametric techniques often rely on assumptions about the shape or parameters of the distribution of time to event. Examples of such technique are the accelerated failure time model. Accelerated failure time models (AFT models) [180] are an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. Babuin et al. [181] determined whether troponin elevations predict in-hospital, short-term, and long-term mortality in medical intensive care unit patients independent of the severity of the underlying disease as measured by the APACHE prognostic system. Wilson et al. used [182] used AFT models to predict cardiovascular risk by using predictors such as age, sex, cholesterol, high-density lipoprotein cholesterol, diabetes mellitus (DM), systolic blood pressure, smoking status, and body mass index (BMI).

Semi-parametric techniques have both parametric and nonparametric components. An example of such a technique is the proportional hazards model. Proportional hazards models relate the time that passes before some event occurs to one or more covariates that may be associated with that quantity of time. In such models, the unique effect of a unit increase in a covariate is multiplicative with respect

to the hazard rate. Yadav et al. [96] used proportional hazards model for risk assessment of comorbid conditions in T2DM. They identified how risks vary across trajectories for the same outcome. The trajectories were defined by using diagnosis codes such as hypertension, hyperlipidemia and T2DM with time to death being modeled as the outcome of interest. Martingale residuals were used to compute the risks. Vinzamury and Reddy [183] extended proportional hazards regression with novel regularization functions to capture correlation and grouping of features effectively. They proposed novel regularization frameworks to handle correlation and sparsity present in EHR data. Further, they demonstrated the applicability of their technique by identifying clinically relevant variables related to heart failure readmission.

5.2 Handling Irregular Time Series Data

Data stored in EHRs is usually collected through longitudinal study. In such studies, the subject outcomes, treatments or exposures are collected at multiple follow-up times, usually at irregular intervals. For example, patients diagnosed with T2DM might be followed over time and annual measures such as Hemoglobin A1c and GFR are collected to characterize the disease burden and health status, respectively. As these repeated measures are correlated within the subject, they require sophisticated analysis techniques. In what follows, we describe techniques that are widely used to handle these repeated measurements. In particular, we cover marginal and conditional models, respectively.

Marginal models are also known as the population averaged model as they make inferences about population averages. In such models, the target of inference is usually the population and these models are used to describe the effect of covariates on the average response. They are also used to contrast the means in sub-populations that share common covariate values. For example, consider a cohort of pre-diabetic patients with elevated cholesterol levels. In this cohort, if we are interested in estimating the progression of patients to full-blown T2DM, we would probably want to use the population-averaged coefficients. Generalized Estimating equations (GEE's) are mostly used for parameter estimation in marginal models. This approach is computationally straightforward and with care can handle missing data, even when the covariance has been misspecified. Generalized estimating equations (GEEs) are used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes. Parameter estimates from the GEEs are consistent even when the covariance structure is misspecified. They are commonly used in large epidemiological studies, especially multi-site cohort studies because they can handle many types of unmeasured dependence between outcomes.

Hernan et al. [184] used marginal models to analyze the causal effect of zidovudine on the survival of human immunodeficiency virus-positive men participating in the Multicenter AIDS Cohort Study. They used a marginal structural Cox model to control further for time-dependent confounding due to CD4 count and other time-dependent covariates and observed a mortality ratio of 0.7. Yu et al. used [185] marginal models to estimate the effect of medication adherence on health outcomes among patients with T2DM. Nandi et al. [186] used them to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes and stroke. King et al. [187] have also discussed the use of marginal models for T2DM related research.

Conditional models [188] are also known as the locally averaged models as they usually make inferences about individual subjects. The estimates are based on averaging or smoothing done by the model, but more locally, are based on sources of dependence in estimating model parameters. For example, consider once again the aforementioned cohort of pre-diabetic patients with elevated cholesterol levels. In this cohort, if we are interested in estimating the effect of statin across every individual, we

would use conditional models.

Yamaoka et al. [189] used conditional models to evaluate the efficacy of lifestyle education for preventing type 2 diabetes in individuals at high risk. They observed that lifestyle education intervention reduced glucose by 0.84 mmol/l as compared to the control group. Mezuk et al. [190] examined the bi-directional prospective relationships between depression and T2DM using a random effects model. Nouwen [191] examined the association of diabetes and the onset of depression by reviewing the literature and conducting a meta-analysis of longitudinal studies on this topic. The conclusion was that people with type 2 diabetes have a 24% increased risk of developing depression.

5.3 Handling Confounding via the Pseudo-outcome Model

In this section, we describe the pseudo-outcome model, which lies at the heart of handling confounding. A confounding variable is an extraneous variable in a statistical model that correlates (directly or inversely) with both the dependent variable and the independent variable. To handle confounding we discuss techniques such as propensity scoring and inverse probability weighing.

Statistical matching techniques such as propensity score matching (PSM) [192] attempt to estimate the effect of a treatment or other intervention by accounting for the covariates that predict receiving the treatment. They aim to reduce the bias caused by confounding variables. PSM creates a group by employing the predicted probability of group membership which is usually obtained from logistic regression. The key advantage of PSM is that by using a linear combination of covariates for a single score, it balances treatment and control groups on a large number of covariates without losing a large number of observations. One disadvantage of PSM is that it only accounts for observed (and observable) covariates. Factors that affect assignment to treatment and outcome but that cannot be observed cannot be accounted for in the matching procedure. Another issue is that PSM requires large samples, with substantial similarities in terms of subjects between treatment and control groups.

Tao et al. [193] used PSM for determining that the costs attributed to type-1 diabetes are disproportionately higher than would be expected given the number of type 1 patients compared with type 2 patients. Austin et al. [194] provided a systematic review of the use of propensity score matching in the cardiovascular surgery literature. Polkinghorne et al. [195] used PSM to analyze the inception and intervention rate of native arteriovenous fistula (AVF). Yasunaga et al. [196] investigated postoperative outcomes after laparoscopic or open distal gastrectomy in Japan One-to-one propensity score matching was performed to compare in-hospital mortality, postoperative complication rates, length of stay, total costs, and 30-day readmission rates between the 2 groups. Short et al. [197] used PSM techniques to examine the effect of beta blockers in the management of chronic obstructive pulmonary disease (COPD), assessing their effect on mortality, hospital admissions, and exacerbations of COPD. They described the additive benefits of beta blockers in reducing oral corticosteroid use and hospital admissions due to respiratory disease. Beta blockers had no deleterious impact on lung function for any treatment step when given in conjunction with either a long acting agonist or antimuscarinic agent. Kuss et al. [198] used PSM to analyze off and on-pump coronary artery bypass grafting techniques. Their review and analysis of propensity score analyses finds off-pump surgery superior to on-pump surgery in all of the assessed short-term outcomes. This advantage was statistically significant and clinically relevant for most outcomes, especially for mortality.

Inverse probability weighting [199] is a statistical technique for calculating statistics standardized to a population different from that in which the data was collected. Instead of adjusting for the propensity score, we could use it to weight the participants. However, there may be prohibitive factors barring

researchers from directly sampling from the target population such as cost, time, or ethical concerns. Robinson et al [**?**] used inverse probability weighting for examining whether lower serum levels of serum 25-hydroxyvitamin are associated with increased risk of developing type 2 diabetes. They used inverse probability weighting to make the study population representative of the WHI population as a whole.

6 Metrics

Quantifying the outcome is the primary interest in many research studies. The outcome is often quantified using various metrics such as incidence rate, prevalence, relative risk and odds ratio [200–203]. Incidence rate [201] indicates the number of new cases of disease in a population at risk over a predefined interval of time. Prevalence indicates the number of existing cases of disease in a population under observation. For example, in a given population of 100,000 persons, there are 980 patients who were diagnosed with tuberculosis within a year and there were 10 patients diagnosed with tuberculosis at a particular point in time. In this scenario, the incidence rate of tuberculosis within a year would be 10/100000 whereas the prevalence rate would be 980/100000.

Relative risk [201] is defined as the frequency of outcome in the exposed group as compared to the frequency of outcome in the unexposed group. For example, consider a cohort of pre-diabetic patients. The cohort is divided into two groups (control and treatment) of 1000 patients each. The treatment group is prescribed statin and the control group is not. The cohort is then followed for 5 years. After 5 years, it was observed that 200 patients in the treatment group progressed to diabetes whereas 100 patients in the control group progressed to diabetes. From this information, the relative risk of diabetes is 2.0: patients within cases are twice as likely to progress to diabetes as controls. Relative risk is 1.0 when the frequency of outcome is same in both the groups. Relative risk greater than 1.0 indicates increased risk of outcome, while less than 1.0 indicates decreased risk (protective effect of exposure).

Odds ratio [202] indicates the odds of exposure/outcome among the case group divided by the odds of the exposure/outcome among controls. For the example above, the odds in the case group will be 0.25 whereas the odds in the control would be 0.10. The odds ratio would then be 25. Similarly, odds ratio can also be defined for cross-sectional, cohort and randomized controlled studies. In the following section, we describe various study designs and metrics which are widely used in respective study designs.

7 Study Design

Study design is the formulation of a hypothesis in medical and clinical research. The aim of a clinical/medical study is to assess the safety, efficacy, and action of a drug or device. They help to examine community, group or national level trends. They are also used to evaluate the effects of multiple exposures and to establish the chronological relationship between exposure and outcome. With the availability of EHR data, a wide variety of clinical studies can be carried out. Clinical studies are categorized based on the underlying design [204]. These study-designs can be primarily classified into two major groups i.e. experimental and observational. Figure 1 succinctly captures this information.

In an **experimental** study design, the researcher intervenes to change the course of the disease and then observes the resultant outcome. Randomized Clinical Trials (RCT's) are examples of experimental study designs. A specific example would be a study where surgery patients with T2DM were

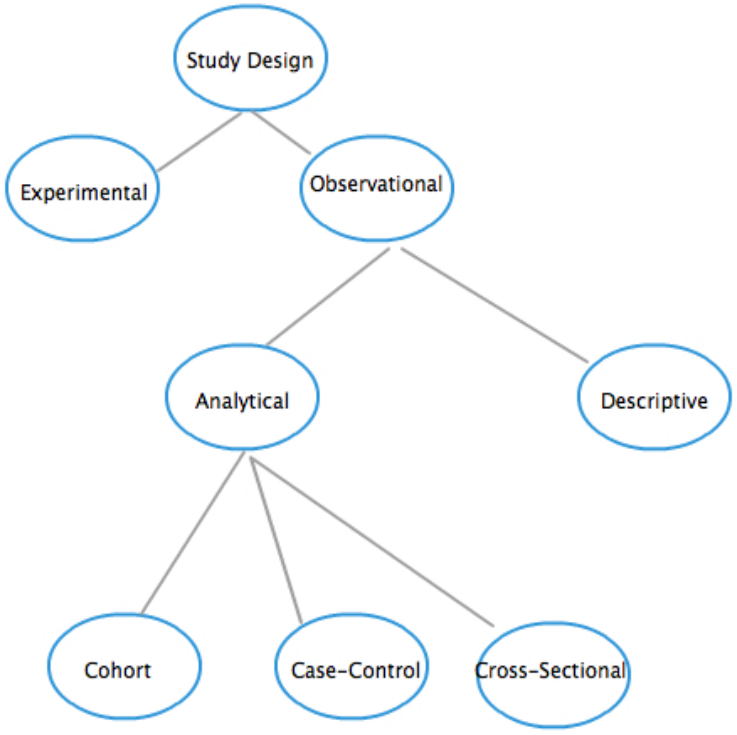

Figure 1: A simple caption

randomized to receive supplemental insulin at bedtime for blood glucose (treatment) or no supplemental insulin (case). As intervention in EHRs is not possible, we will not discuss these study designs in great detail.

By **observational** [205], we refer to study designs where the researchers do not intervene. In such studies, the investigators observe subjects and measure variables of interest without assigning treatments to the subjects. The treatment that each subject receives is beyond the control of the investigator. For example, consider a study that investigates the effect of smoking(exposure) on lung capacity (outcome). A cohort of young men aged 18-25 are recruited. Some subjects in this cohort smoke tobacco (exposed group) and some do not (exposed/comparison group). The investigator has no ability to influence the exposure since the subjects smoking behavior is uninfluenced by the investigator. This cohort is then followed for a number of years to analyze the effect of smoking on lung capacity by comparing the exposed group with the unexposed group [206, 207]. Observational studies can be further categorized as analytical (if there is a comparison group (i.e. case and control)) or descriptive (no comparison group).

Analytical studies are mostly used to test hypotheses by selection and comparison of groups. They also aim to identify risk and protective factors for diseases as well as causative associations between exposures and outcomes. Analytical studies [201] can be further divided into three major groups based on the temporal direction in the study. Studies which start with an outcome and look back in time for exposure are known as **retrospective studies**. If the study begins with an exposure and concludes with an outcome, we refer to them as **cohort studies** [208–211]. It involves following subjects over time to analyze the effect of exposure. If we only consider a single point in time, where the outcome and the exposure are both present at that time, we refer to the study as **cross-sectional** [201] . Such studies mostly involve the selection of a sample of the population, irrespective of the outcome and the

exposure. Alternatively, these studies might represent a snap-shot of the underlying patient population.

Descriptive study designs mostly deal with the frequency and the distribution of risk factors in populations and enable us to assess the extent of a disease of interest. These study designs are usually used to build hypotheses, thereby building the framework for future clinical research.

In the following section, we will discuss the aforementioned study-designs along with clinically relevant examples. We will also discuss how certain studies might incorporate biases and confounding factors.

7.1 Retrospective Studies

Retrospective studies are study designs which look backwards, i.e., the study groups are defined using an outcome and the study looks back in time to analyze the exposure status of a subject. They are often used to identify risk factors that may contribute to a medical condition by comparing subjects who have that condition/disease with patients who do not have the condition/disease but are otherwise similar [212–216].

These study designs are very useful in the investigation of diseases that have a long latency period, such as cancer and diabetes as cohort studies (discussed next) involve many years of follow-up before the outcome becomes apparent. Since such studies have treatment and control identified right at the beginning of the study, they are very efficient in terms of time and effort.

However, when the exposure rate is low these study designs are inefficient as researchers would have to examine many cases and controls to find one patient who had exposure. For example using a case-control study design to investigate the effect of pancreatic cancer (exposure) on T2DM (outcome) would be impractical because the exposure is very rare. When the exposure rate is low, cohort studies should be the default standard. Moreover, choosing a control group and obtaining exposure history might greatly affect a study's vulnerability to bias. Improper selection of the control group can also bias the results of the study and therefore researchers should provide clear eligibility criteria for the outcome being studied, such as age, gender, racial makeup and ethnicity. These studies often come under the realm of temporal supervised learning techniques.

In risk prediction, case control study designs are widely used due to their ability to expose the association between risk factors (exposure) and outcome. Consider for example a cohort of diabetic patients (case) and non-diabetic patients (control). We track these patients backwards in time for a fixed number of years (i.e. baseline) to explore the exposure. At baseline, we investigate whether the patients in the case and the controls were obese (exposure). Using the patients' baseline characteristics as exposures we can determine the patients' odds of progressing to T2DM if the patient is obese. Since we followed the patients from outcome to exposure, we can estimate the odds ratio i.e. the proportion of individuals exposed in each of the case and the control group.

7.2 Cohort Studies

Cohort studies are also known as incidence, longitudinal, forward-looking, follow-up, concurrent or prospective studies [217]. In such studies we compare the experience of a group exposed to some intervention with another group not exposed to the same intervention. The underlying characteristic of such studies is that they track people forward in time from exposure to outcome [218–220]. As an

illustration, consider a group of patients, some diagnosed with obesity at a particular point in time. Our interest is to investigate the relationship between obesity and diabetes. This population contains patients with exposure (i.e. obese patients), outcome (diabetes) and both (obese diabetic patients). We then exclude all diabetic patients and only retain patients without diabetes; this is our study cohort. Some of patients in the cohort have the exposure (obese) and others do not. We follow the cohort forward in time and observe how many patients convert to diabetes (cases) and how many remain non-diabetic (controls) both among the exposed and among the unexposed patients. This design ensures that exposure precedes outcome thus it allows us to estimate incident rates and the relative risk (or odds) of incident diabetes.

Cohort studies are considered to be the best study designs for ascertaining both the incidence and natural history of a disorder as the temporal sequence between the cause and the outcome is usually clear [169]. They are also useful in analyzing multiple outcomes that might arise after a single exposure. For example, smoking (i.e. exposure) might lead to multiple outcomes such as stroke, oral cancer and heart disease. They are often utilized to explore rare disease phenomenon. For example, to investigate the effects of ionizing radiation in the workplace, subjects might be selected from factories or hospitals thereby avoiding the ethical issues arising due to exposure assignment.

However, such study designs come with certain caveats. Firstly, selection bias is inherent in such cohort studies [221]. For example, in a cohort study analyzing effects of smoking on T2DM, those who smoke would differ in other important ways (lifestyle) from those who do not smoke. In order to validate the effect of exposure (i.e. smoking), both the (case and controls) must be similar in all respects except for the absence/presence of exposure and the outcomes. Secondly, loss of subjects due to censoring can be a difficulty, even when the study is short, but particularly with longitudinal studies that continue for decades. For example, progression from T2DM to associated complications such as PVD and IHD takes around 5 to 10 years, and subjects may drop out over this long period.

In risk prediction, cohort study designs are widely used due to their ability to expose the association between risk factors (exposure) and outcome. Consider for example a cohort of pre-diabetic patients, many of whom have different conditions diabetes (e.g. obesity, high cholesterol, high blood pressure) comorbid with T2DM at baseline. We follow this cohort for a number of years. Some of the patients progress to overt diabetes (case) and others remain non-diabetic until the end of the study (control). Some patients are lost to follow-up before the end of the study (censored). Now, we can consider the patients' baseline characteristics (e.g. obesity, high cholesterol or high blood pressure) as exposures and determine which of these exposures increase (or decrease) the patients' risk of developing diabetes significantly. Beside risk prediction, this application also relates to biomarker (risk factor) discovery.

Cohort design can also be valuable for subpopulation mining. In the above example, we can examine the effect of simultaneously being obese, having high cholesterol *and* high blood pressure at baseline on incident diabetes. The set of conditions (obesity, high cholesterol and high blood pressure) define a subpopulation which we can consider as an exposure. Thus the cohort design can be used for subpopulation mining.

7.3 Cross-Section Studies

Cross-sectional studies fall under the category of analytical studies, which are characterized by seeking a comparison between cases and controls by collecting data at one specific point in time - that is the cross-sectional data [222]. Such study designs differ from retrospective and cohort studies in that they

aim to make inferences based on data that is collected only once rather than collected multiple times [223].

Cross sectional studies are frequently used to analyze the presence or absence of a disease and outcome at a particular point of time across the case and the control group. They are mostly used to investigate the association between the risk factor and the outcome [224–226]. Due to this, metrics such as prevalence are widely used in these study designs [222]. For example, researchers might measure and compare the cholesterol levels of two age groups - over 40 and under 40 for joggers, and compare these to cholesterol levels among non-joggers in the same age groups. Researchers might even create subgroups for gender. Thus cross-sectional study designs allow researchers to compare many different variables simultaneously.

Some patients in the population at the time of the study will have the exposure, some will have the outcome and others will have both [227]. Since it is a single point in time, the temporal relationship between exposure and outcome cannot be determined. From the proportions of patients with exposure, outcome and both, we can estimate the relative odds of outcome given exposure. We can also estimate the prevalence of outcome but not the incidence rate of outcome. To illustrate, in our last example, we cannot know for sure if our joggers had low cholesterol levels before taking up their exercise regimes, or if the behavior of daily jogging helped reduce cholesterol levels that had previously been high. Similarly, we would not compare past or future cholesterol levels for both the groups, for these would fall outside the frame. We would look only at cholesterol levels at one point in time.

In risk prediction, cross-sectional study designs are widely used due to their ability to expose the association between risk factors (exposure) and outcome using data at a single point in time. Consider our old cohort of prediabetic patients, many of whom have different conditions (e.g. obesity, high cholesterol, high blood pressure) at baseline. Some of the patients might have obesity (case) while others are healthy (control). Now, we can compare the groups for other outcomes of interest (e.g. high cholesterol or high blood pressure) with respect to our exposure (obesity).

7.4 Descriptive Studies

Descriptive studies are designed to describe the existing distribution of variables, without regard to causal or other hypotheses [169, 200]. In such studies, apart from age and gender, other characteristics such as race, occupation and recreational activities are often described [228–234]. Descriptive studies are often classified into multiple categories based on whether they deal with individuals or populations. For example, studies reporting an unusual disease or association or surveillance studies over a community are examples of descriptive studies based on individuals. Examples of descriptive studies based on populations can be correlational studies looking for associations between exposures and outcomes. Correlational studies often lead to hypotheses for more advanced study designs. These studies often come under the realm of non-temporal unsupervised learning techniques. The defining characteristic is that there are no cases or controls as compared to cohort, cross-sectional and case-control studies and hence no comparisons.

Descriptive studies are often useful for analyzing the medical state of population and health-care planning [235–238]. For example, such study designs are widely used to investigate the tobacco consumption within a population, age group, gender or socio-economic class.

Such study designs do not provide us with the platform to carry out temporal reasoning and causal

research. Since there are no comparison groups, no inference can be derived from the cases and the controls.

In comorbidity analysis, descriptive study designs are widely used due to their ability to expose the distribution of diseases within any population of interest. For example, consider our cohort of pre-diabetic patients, many of whom have different conditions (e.g. obesity, high cholesterol, high blood pressure) at baseline. Using such study designs, we can estimate the prevalence of these comorbid conditions. Further such analysis can lead to future estimation of sequential patterns in which such diseases occur.

8 Methodology

The discipline of EHR data mining stands at the intersection of epidemiology, biostatistics and general data mining. From epidemiology and biostatistics, we have borrowed study design, the methodology that allows us to organize our EHR data into a matrix that is amenable to the application of data mining algorithms that can correctly answer meaningful clinical questions. We have also borrowed basic approaches to address the challenges that EHR data posed including censoring, analysis of irregular time series data and methodologies for causal inference[1]. In this section, we focus on the contributions of general data mining.

Traditionally, data mining techniques are broadly categorized as supervised or unsupervised: supervised methods take an outcome into account, while unsupervised methods simply learn from the structure of the data. The hallmark of EHR data is its temporal nature, suggesting that data mining techniques be further categorized based on their ability to take time into account. We call a data mining algorithm and its resulting model *time-aware*, if its output depends on time; and we call it *time-agnostic*, if it builds a model that does not take time into account.

Although EHR data is inherently temporal, time is not always of relevance. The clinical question we aim to answer may be *temporal* if time is of relevance (i.e. time is part of the question) or it may be *atemporal* (not temporal) if time is not part of the question. Atemporal questions are naturally answered by time-agnostic data mining techniques. On the other hand, temporal questions can be either answered by time-aware models or if the question can be transformed into a simpler atemporal question, it can also be solved using time-agnostic models. For example, predicting the risk of 30-day mortality after surgery is a temporal question (time is part of the question) but it can be solved using time-aware models (e.g. Cox model) or time-agnostic models (e.g. logistic regression).

The study design dictates whether a question can be temporal or atemporal and it also determines in large part whether any of the challenges posed by the EHR data can be successfully addressed. For this reason, we describe data mining techniques that are commonly applied in the context of the applicable study designs.

8.1 Descriptive Studies

Descriptive studies represent the broadest variety of inquires we can undertake, ranging from simple statistics (prevalence rate, incidence rate) to descriptions of the progression of a particular diseases via

[1]The roots of causal inference are in computer science, but has been embraced by epidemiology and biostatistics resulting in the development of the advanced techniques we described earlier.

case studies. Such simple applications do not require data mining, but data mining techniques enable more advanced applications including comorbidity analysis and trajectory mining. While descriptive studies cover a wide range of applications, their defining characteristic is that no comparison is made between patients (or patient groups) with and without a particular outcome. Without a particular outcome, we cannot have outcome labels hence the problem at hand is *unsupervised*.

Descriptive studies are commonly utilized to answer both temporal and atemporal clinical questions. For example, estimating prevalence rates at a particular time is an atemporal clinical question, while extracting the trajectory of a patient as sequences of diagnosis codes is naturally a temporal clinical question. Therefore both time-aware and time-agnostic data mining techniques are applicable to descriptive studies.

8.1.1 Atemporal Descriptive Studies

Atemporal descriptive techniques are arguably the simplest methods, typically textbook methods. A prototypical application of this nature would be to take a cross-section of the population at a particular time and cluster the patients based on the conditions they present. Textbook data mining techniques have a limited ability to handle the temporal aspect of EHR data. One option is to use specialized techniques, such as sequence mining, while another is to "flatten" the temporal dimension of the data through temporal abstraction, e.g. by applying features and apply non-temporal unsupervised techniques.

Unsupervised techniques applied to non-temporal data have been widely used for *identifying clusters* of patients that have similar characteristics (e.g. demographics, medications, diagnosis codes, laboratory test results) and for finding *associations* between clinical concepts (e.g. medications, diagnosis codes and demographic attributes).

Clustering: Gotz et al. [239] used clustering techniques for identifying a cohort of patients similar to a patient under observation. They used the cohort as a surrogate for near-term physiological assessment of the target patient. Roque et al. [43] stratified patients using hierarchical clustering, where the distance between patient records was computed using the cosine similarity of diagnosis codes. Roque et al. further used this stratification for comorbidity analysis. Bauer-Mehren et al. [240] used medical concepts (medication information, diagnosis codes, procedure codes) for patient stratification, where the Jaccard index was used as the similarity measure. Along similar lines, Doshi et al. [44] investigated the patterns of co-occurring diseases for patients diagnosed with autism spectrum disorders (ASD). They identified multiple ASD related patterns using hierarchical clustering. They further discussed how the aforementioned patterns can be attributed to genetic and environmental factors. Kalankesh et al. [241] noted that representing the medical state of a patient with diagnosis codes can lead to sparse clusters. To overcome this problem, they used Principal Component Analysis (PCA) [242] to reduce the dimensionality, thereby making the structure more amenable for visualization and clustering. Marlin et al [243] developed a probabilistic clustering method to mitigate the effects of temporal sparsity inherent in EHRs. They used unsupervised learning techniques for automatically uncovering insightful patterns from physiologic time-series data.

Association Analysis: Association rule mining techniques [244] such as Apriori have also been used on EHR data to identify associations among clinical concepts (medications, laboratory results and problem diagnoses). Wright et al. [245] used the Apriori framework to detect transitive associations between laboratory test results and diagnosis codes and between laboratory test results and medications.

For example, they observed some unexpected associations between hypertension and insulin. They attributed this finding to co-occurring diseases and proposed a novel way to identify such transitive associations. Cao et al. [46] used co-occurrence statistics to identify direct and indirect associations among medical concepts. Holmes et al. [47] used statistical approaches to detect associations between rare diseases. They observed that analyzing cohorts comprised of sick patients leads to identification of significant findings. Shin et al. [48] used association rule mining to identify co-morbidities (e.g. non-insulin dependent diabetes mellitus (NIDDM) and cerebral infarction) which are strongly associated with hypertension. Hanauer et al. [246] used statistical tests to observe common pathways for diseases such as granuloma annulare and osteoarthritis.

As these studies often deal with information collected at one time instant or summarized until the time of interest, causal analysis is not feasible. Further, as these study designs lack a case and a control group, identifying causal factors leading to the outcome of interest is a challenging problem. However associating predictors with outcomes can be carried out with ease. In order words, it is not possible to distinguish whether the outcome of interest (i.e. condition) preceded or followed the condition. Although the study design is inherently atemporal, longitudinal data can still be used for research purposes. In such studies longitudinal data is often summarized or aggregated. Examples of summarization and aggregation include computing the mean, median, averages, variance, higher order moments and shaplets using temporal logic rules. On one hand, these aggregation techniques convey meaningful information about temporal and seasonal trends, but on the other hand they are highly susceptible to outliers and noise.

How to handle censored data is often an issue in such study designs. Data incorporating right censoring cannot be used for modeling as such patient records often have no information about the outcome variable. Therefore, there is no way to ascertain the prevalence of existing conditions with outcome variables. However, patient records susceptible to left censoring might not be discarded because we may have no information for some characteristics, we can still model them as unknown quantities. Typically this would require sophisticated research techniques. Further, in clustering, the challenges lies in the semantic differences between the groupings as patient stratification using ICD codes might lead to biases. Such biases arise because ICD codes are often generated for billing related purposes.

8.1.2 Temporal Descriptive Studies

Time plays an important role in the clinical questions. For example, the sequence of events, timing between events, etc. Standard textbook data mining techniques exist to solve such problems (e.g. sequence mining, Markov models, etc), but to achieve better results, significant improvements have been proposed. We broadly classify the approaches that can be carried out using such techniques as those which use time-aware techniques, e.g., sequence mining, time-lagged correlations, etc., and those which simplify the problem and apply time-agnosic techniques e.g., temporal-abstraction (summarizing the longitudinal data) and HMM trajectory clustering (usingHMM to simplify away time so that standard clustering is applicable).

Sequential Rule Mining: Researchers have explored sequential association rule mining techniques for identifying causal relationships between diagnosis codes. Hanauer and Ramakrishnan [247] identified strongly associated pairs of ICD-9 codes with varying numbers of strong temporal associations ranging from 1 day to 10 years apart. They observed interesting temporal relationships between hypothyroidism and shingles (herpes reactivation). Liao and Chen [248] proposed a sequential pattern mining approach to mine sequences with a gap constraints. Such gaps represent the delay between

two concepts. Hripsack et al. [249] measured lagged linear correlation between EHR variables and healthcare process events. In their analysis, they considered five common healthcare process events: inpatient admission, inpatient discharge, outpatient visit, emergency department visit and ambulatory surgery and computed their correlation with several EHR variables such as laboratory values and concepts extracted from clinical notes.

Temporal Abstraction Framework: The temporal abstraction framework has been frequently used to extract patterns from EHR data. Patterns can be abstracted using state representations(e.g. high, medium or low) or trend representations(e.g. increasing, decreasing, constant). Shahar et al. [250] provided a mechanism to abstract patterns from unevenly spaced time-series. Such time-series are common in EHR data elements such as laboratories test results and vitals. They further proposed temporal logic relations to combine patterns generated from univariate time-series. Sacchi et al. [251] extended the temporal abstraction framework to generate temporal association rules (TARs). In TAR's, the antecedent and the consequent both consist of temporal patterns generated using the temporal abstraction framework. Jin et al. [252] further extended the TAR framework, to generate rules for mining unanticipated episodes where certain event patterns unexpectedly lead to outcomes e.g. taking two medicines together sometimes causes an adverse reaction. Batal et al. [253] used the temporal abstraction framework to propose the Segmented Time Series Feature mining algorithm for identifying the frequent patterns from an unevenly sampled time-series. Such modeling techniques have their own set of challenges. Patterns generated from individual patient time series are susceptible to noise. Further, such patterns can be of uneven temporal duration.

Dynamic Clustering: Clustering techniques have also been used to group EHR data. Ghassempour et al. [254] used hidden Markov models (HMM) to cluster patient medical trajectories. In their approach, they used both categorical variables (diagnosis codes) and continuous variables (vitals and laboratories test results) for clustering. They first mapped each medical trajectory to an HMM and then used KL divergence to compute the distance between two HMM's.

Visualization: Research has also been carried out in analytical reasoning facilitated by advanced interactive visual interfaces. Several research has been carried out by highlighting the opportunities and associated challenges [339], cohort analysis and exploration [340, 341], exploring comorbidities [342, 343], exploring concepts [344], clinical decision support [345], cohort identification [346], disease network visualization [347] and temporal frequent event sequences [348].

As there are no comparison groups present (i.e. no case and control) any exploration for causation of disease or outcome of interest is not possible. In this aspect, they are highly reminiscent of atemporal descriptive study designs. The major difference with atemporal descriptive studies being that for atemporal descriptive studies, only prevalence rates can be computed, while for temporal descriptive studies, prevalence as well as incidence rates can be computed. Another advantage of such studies lies in their ability to handle censored data. All kinds of censored data such as right censored, left censored or interval censored can be managed using such techniques. For example, patient records with certain clinical characteristics but no information about the outcome of interest can be used for modeling purposes. For such analysis, survival regression methods are often used. These methods mostly aim at modeling the time to event data.

In these studies, data for every patient is often available at multiple time instances. With availability of this data, sophisticated and rigorous techniques that can model time-varying covariates and outcomes can be efficiently utilized. However, until now not much research has been carried along these lines. In the past, research has focused more on lagged correlations and sequential patterns. Al-

though the aforementioned approaches are quite informative, they are also quite susceptible to biases and confounding effects.

8.2 Cross-Sectional Design

Cross-sectional studies are carried out by collecting data at one time point. The aim of such studies is usually to estimate the prevalence of the outcome of interest i.e. to investigate the associations between risk factors and the outcome of interest. In such studies, data is often collected on individual characteristics, such as exposure to risk factors, demographic attributes and information about the outcome. In what follows, we will describe the techniques often used for such study designs along with examples of research carried out in the past.

When a study is designed as cross-sectional, supervised non-temporal data mining techniques are the natural modeling choices. When the study is inherently temporal and it employs a case-control or cohort design, it can still be solved using supervised non-temporal techniques, but we incur some loss of information. Supervised non-temporal techniques allow for having a well-defined outcome but have no facility to extract the temporal information from the data. In other words, the studies described in this section may be temporal in nature, but the algorithms that were used to solve them are non-temporal. For example, a study investigating the 30-day mortality of patients following an exposure can be modeled using supervised non-temporal techniques as long as we only consider a binary outcome, namely, whether the patients survived for 30 days or not. If our primary interest is the time itself, and we wish to model the length of time during which the patients actually survived we would have to employ supervised temporal techniques. Analogously, transforming time-dependent predictors to non-temporal predictors through temporal abstraction is possible, allowing for the application of supervised non-temporal techniques to complex temporal study designs—naturally, at the cost of losing information. Since interest in a specific outcome is very natural and there is great appeal in simplifying these problems to become solvable through relatively simple supervised non-temporal data mining techniques, such techniques have been applied to a broad spectrum of problems, including risk prediction for hospitalization, re-hospitalization, diagnostic and prognostic reasoning.

Rule Based Methodologies: White et al. [138] conducted a large scale study for analyzing web search logs for detection of adverse events related to the drug pair, paroxetine and pravastatin. They analyzed whether the drug interaction leads to hyperglycemia. Iyer et al. [136] used NLP techniques for mining clinical notes to identify events related to adverse drug-drug associations. They believed that EHRs contain rich information in the unstructured notes. Haerian et al. [137] hypothesized that adverse events might be caused by the patient's underlying medical condition. Along similar lines, Vilar et al. [134] used disproportionality based techniques to analyze adverse drug events related to pancreatitis, Li et al.[255] used penalized logistic regression to analyze associations between ADRs and Epstein et al. [135] used NLP techniques to analyze medication and food allergies. Supervised non-temporal methodologies have been frequently used in the form of rule-based techniques for cohort identification. Phenotyping algorithms for diseases such as celiac disease, neuropsychiatric disorders, drug-induced liver injury, and T2DM [256–258] have been widely explored. Supervised pattern mining approaches using the temporal abstraction framework have been used for predicting Heparin Induced Thrombocytopenia (HIT) [253]. Batal and Hauskrecht [259] used such methodologies to generate minimal predictive rules for Heparin Platelet Factor 4 antibody (HPF4) test orders. They further extended their approach by introducing the minimal predictive patterns (MPP) framework wherein they directly mine a set of highly discriminative patterns [260]. Those patterns were later used for classification related tasks.

Bayesian Networks: Bayesian Networks have also been used to model EHRs for diagnostic reasoning (constructing the medical state of the patient using laboratory test results), prognostic reasoning (prediction about the future), and discovering functional static interactions between the outcome and the predictors [261]. Zhao et al. [97] integrated knowledge from Pubmed along with EHR data to develop a weighted bayesian network for pancreatic cancer prediction. They also discussed how their approach can be used to detect clinically irrelevant variables for disease prediction. Sverchkov et al. [262] compared clinical datasets by capturing the clinical relationships between the individual datasets by using the Bayesian networks. The multivariate probability distributions were then used to compare the clinical datasets.

Numerous issues and challenges arise when we analyze EHR data using such study designs. The foremost issue is causality. It is limited by the nature of such study designs, as information is usually collected at one time point and hence, it gives no indication of the sequence of events: whether exposure occurred before, after or during the onset of the disease outcome. Inferring causation with this caveat might lead to erroneous findings and thus it is impossible to infer causality. By virtue of their design, longitudinal analysis is not possible in such studies. However, techniques (as mentioned in atemporal descriptive study designs) such as the temporal abstraction framework or qualitative abstraction techniques such as by computing the mean, median, mode, variance or slope are widely used to employ time-agnostic strategies. The substantial difference go atemporal descriptive study designs is the availability of comparison groups in these studies. They provide a platform amenable for applications such as adverse event detection, and cohort identification. In terms of censoring, right censored data poses substantial challenges, as no information about the outcome is present. However, left censored and interval censored data can be handled by such techniques to a great extent.

8.3 Cohort and Retrospective Study Design

Cohort and Retrospective studies compare patients groups with different exposures over time and record their outcomes. They differ in the direction in which time is observed: in cohort studies patients are followed from exposure to outcome and in retrospective studies, patients are followed from outcome to exposures. While this difference has far-reaching consequences on the required sample sizes, exposure rates and the metrics we can estimate, once the design matrix has been constructed, the same data mining methods apply to both of these study designs. Hence we consider these two designs together.

What is common across these study designs is that they are best suited to answer *temporal* questions; if time is not of interest, a cross-sectional study would suffice. As it is typical with temporal questions, we can use either time-aware models or we can simplify the question such that it can be answered using time-agnostic models. In the following paragraphs, we provide examples of both.

Time-Agnostic Models for Cohort and Retrospective Studies

Time Agnostic Regression: Supervised time-agnostic models are commonly employed when time-to-event can be removed from the clinical question. For example, time-to-rehospitalization can be simplifed to the binary outcome of 30-day rehospitlaization (yes/no) of 30-day-rehospitalization (yes/no) which does not include time. Applications of supervised time-agnostic modeling include predicting the onset of neonatal sepsis [263], potentially preventable events [264], 30 day hospital readmissions [265, 266], post-hospitalization VTE risk [267], [268], T2DM risk forecasting [269], atrial fibrillation [270], 5 year long life expectancy risk calculation [271], risk of depression using diagnosis codes [272], survival of heart-lung transplant patients [273], breast cancer survivability [274], 30 day

mortality in patients suffering with cardio-vascular diseases, risk of retinopathy in patients suffering from type 1 diabetes mellitus (T1DM) [275], mortality in patients suffering from acute kidney injury [276], mortality prediction in ICU [277] and risk of dementia [278]. For these analyses, almost all flavors of common predictive modeling techniques (decision trees [269, 274] , [279], ensemble techniques (e.g. bagging, boosting, random forests) [263, 265, 267], [270], nave Bayes [267, 270, 274], linear regression, support vector machines[269] and logistic regression [265, 268, 269, 272, 280] have been used. These techniques have also been used for identification of regional differences in breast cancer survival rates despite guidelines [281], comparison of cancer survival rates across continents [282], comparison of cancer and survival patients over time, exploring relationships between hospital surgical volumes and 5 year relationship of stomach cancers [283], comparing dosage volumes of warfarin in European-American and African-American [284], postpartum depression rates in Asian-American subgroups (Indian, Chinese, Filipino, Japanese, Korean, Vietnamese) [285], analyzing the effect of different ethnicities on different levels of susceptibility to diabetes related complications and studying the detrimental effect of fibrates on women as compared to men in a population presenting with high cholesterol levels.

Ghalwash et al. [286] proposed data-driven predictive models to find a suitable duration of the hemoadsorption (HA) therapy control and observed that their method applies the therapy in non-continuous fashion, which results in substantial monetary savings. Sun et al. [287] worked on predicting the risk and timing of deterioration in hypertension control by analyzing the transition points at which hypertension is brought into as well as pushed out of control. Wang et al. [288] developed a dynamic Poisson autoregressive model with exogenous input variables for flu forecasting where in they allowed the autoregressive model to change over time. Panahiazar et al. [289] built a heart failure risk prediction model using several machine learning techniques where in they included multiple comorbidities which lead to improvement in prognostic predictive accuracy. Wang et al. [290] proposed Multilinear sparse logistic regression to handle data in the form of multi-dimensional arrays. They used their methods to predict the onset risk of patients with Alzheimer's risk and heart failure.

Such techniques also have their own share of caveats. Causal analysis is not possible as time-to-event data is often ruled out and there is no way to ascertain the relationship between diseases and the outcome of interest. The inherent design of such techniques rules out longitudinal analysis. Temporal abstraction is also employed to summarize time. As comparison groups are available in such study designs they are well-suited for applications such as risk prediction. Further, handling right censored is not possible but handling left censored data and interval censored data is plausible.

Time-Aware Models for Cohort and Case/Control Studies

Supervised time-aware models are utilized when the clinical question cannot be simplified or if the simplification to time-agnostic modeling comes at a significant loss of information. Such question focus on the time-to-event itself (clearly cannot be simplified), sequences of events or when time-to-event carries additional information about the outcome. Continuing with the example of 30-day rehospitalization, by simplifying the outcome to binary yes/no, we lose information since we ignore whether the patient was re-hospitalized in (say) 7 days vs 20 days. The former case is clearly more severe.

Many of the temporal clinical questions are related to right censoring. Survival modeling, which was specifically developed for this purpose, is the quintessential technique for this study design. Survival modeling is a suite of techniques with various specializations that share a common characteristic of being able to handle time and censoring.

Survival Modeling: Wells et al. [291] hypothesized that patients diagnosed with T2DM have an increased risk of mortality. They used Cox proportional hazards regression with time to death as the outcome. They also observed that certain interaction terms involving medications and age were significant indicators. Vinzamury and Reddy [292] extended Cox proportional hazards regression with novel regularization functions to capture correlation and grouping of features effectively. They proposed novel regularization frameworks to handle the correlation and sparsity present in EHR data. They demonstrated the applicability of their technique by identifying clinically relevant variables related to heart failure readmission. Vinzamury et al. [183] proposed a novel active learning based survival model wherein continuous feedback from a domain expert can be utilized to refine the model. Survival modeling techniques on time-to-event data have been explored widely in the past. Cox regression [175, 293] is one of the most commonly used survival regression models. Its formulation, namely its semi-parametric nature, with the mild assumption of the proportionality of hazards, makes it ideal for many practical applications in fields such as economics [294], healthcare [295–297] and recommendation systems [298].

Cox models, as most other regression techniques, are susceptible to overfitting. Standard regularization techniques, developed for other regression methods, have been applied to Cox models, as well. Lasso [299] and elastic-net regularized Cox models [300] have been developed, and have been further extended by regularizing them with convex combinations of L1 and L2 penalties [301]. We are not aware of regularization for time-dependent covariate Cox models [302], which would be a straightforward extension.

Reddy et al. [183] proposed an active learning based survival model which uses a novel model discriminative gradient based sampling scheme and observed better sampling rates as compared to other sampling strategies. They also proposed correlation based regularizers with Cox regression to handle correlated and grouped features which are commonly seen in many practical problems [292]. Similarly Gopakumar et al. proposed a stabilized sparse Cox model of time-to-events using clinical structures inherent in Electronic Medical Records. They estimated the feature graph derived from two types of EMR structures: the temporal structure of the disease and intervention recurrences, and the hierarchical structure of medical knowledge and practices [303]. To handle the high-dimensionality of high-throughput genomic data, Kuang et al. [304] extended Cox models by proposing network-based Cox regression model called Net-Cox and applied Net-Cox for a large-scale survival analysis across multiple ovarian cancer datasets.

Support vector machine [305] models have also been extended to handle censored data [306–310]. In such techniques, often the task is converted into a ranking problem via the concordance index. This in turn is efficiently solved using convex optimization techniques. Along similar lines, Khosla et al. [311] proposed a margin based censored regression algorithm which combines margin-based classifiers with censored regression algorithms to achieve a better concordance index. They used their technique to identify potential novel risk markers for cardiac problems.

Research has also been carried out on extending decision trees to handle censored data [312]. Ishwaran et al. [313] proposed Random Survival Forests for analyzing right censored survival data. They analyzed splitting rules for growing survival trees, introduced a new measure of mortality and applied it for patients diagnosed with coronary artery disease. Neural nets have also been adapted to handle censored data with varying results [314, 315]. Techniques such as reverse survival [316] have also been explored in the past wherein they go further back in time.

Dynamic Bayes Networks: While survival models are by far the predominant type of mod-

els, other methods that can incorporate temporal information also exist. Dynamic Bayesian networks (DBN) have been used to model temporal relationships among EHR variables [317]. Nachimuthu et al. [144] used DBN's to model temporal relationships between insulin and glucose homeostasis. The modeling was further used to predict the future glucose levels of a patient admitted in an ICU. They also discussed the reasons for using first-order Markov models to model the temporal relationships. Sandri et al. [318] used DBNs with multiple order dependencies to impose restrictions on the causal structure, while modeling organ failure in patients admitted to an ICU. In their model, each time-stamp represented a day. They further imposed several constraints such as that no patient discharges were recorded on the second day and that all patients were either deceased or considered discharged on their seventh day. Such constraints were imposed to reduce complexity of the model. Along similar lines, Rose et al. [145] used DBN's to assist physicians in monitoring the weight of patients suffering from chronic renal failure, Gatti et al. [319] used it to model heart failure and Peelen et al. [320] used hierarchical DBN's for modeling organ failure. Expectation-Maximization was used to learn conditional probabilities in these DBN's.

In the realm of supervised temporal pattern mining, research has extended the temporal abstraction framework by mining recent temporal patterns for monitoring and event detection problems in patients suffering from diabetes [260]. Sengupta et al. [321] used similar techniques for detecting sequential rules associated with the early identification of brain tumors. Simon et. al. [92] proposed survival association rule mining (SARM) techniques which uses survival modeling techniques to incorporate the effects of dosage and other confounders such as age and gender.

These techniques are by far the most successful in terms of overcoming EHRs related challenges. Right, left and interval based censoring can be easily handled by employing techniques such as Cox proportional hazards regression and accelerated failure models.

The biggest claim of such techniques is their ability to handle causation. As these techniques have comparison groups (i.e. case and control) and can handle time-to-event data, causal analysis can be performed with ease. Further, causation by adjusting for measured confounders can also be analyzed by using marginal structural models and structured nested models. However the literature of such techniques in computer science is very sparse. One area, where more work should be done is to handle unmeasured confounders for the disease of interest. Similarly more research needs to be focused in areas where the effects of confounders need to be adjusted for time-to-event data.

Applications	Descriptive		Cross Sectional	Retrospective or Case-Co	
	Atemporal	Temporal		Time Agnostic	Time
Understanding the Natural History of Disease	[46, 47, 242, 244, 246]	[247–249, 251–253]	[49]		[144, 260
Cohort Identification	[43, 240]	[65, 72]	[66–68, 323]	[257, 258, 274]	[324]
Risk Prediction/Biomarker Discovery	[90, 239]		[79–84]	[97, 183, 263, 276, 278, 287, 288, 291, 325–331]	[319]
Predicting the next complication: What and When				[273, 275, 286, 332]	[292, 318
Quantifying the effect of Intervention				[91]	
Patient Medical Trajectories	[88]	[254]			[96]
Constructing evidence based guidelines		[132]			
Adverse Event Detection			[136–138, 140, 143, 256–258]		

Table 1 provides a succinct representation of the major work done using EHRs. The rows correspond to the major application areas, which we broadly discussed in Section 2. The columns represent the methodologies categorized into groups, which we presented in section 8. Building on this understanding, we will explore, discuss and present novel insights about how data mining techniques have been utilized for EHRs. In particular, we analyze why certain areas of EHRs are widely popular, why others are virtually unexplored, and try to identify areas that appear ripe for new research.

A quick glance at Table 1 reveals that substantially more work has been carried out in supervised settings as compared to the unsupervised setting. This difference is not accidental, but rather stems from the nature of research in the medical domain, as research in medical sciences has hitherto been driven by pre-defined outcomes. Corroborating this fact, every randomized clinical trial initially has a well-defined clinical question. Conversely, research in unsupervised domains often leads to the discovery of redundant, widely known facts. The high dimensionality and associated heterogeneity of EHR data lead to increased complexity thereby very large amounts of data is required to discover meaningful relationships through unsupervised techniques.

Another observation from Table 1 is that risk assessment has been widely explored in the health care industry. In risk analyses, the goal is to compute the probability of a patient's progression to an outcome (e.g. diabetes) of interest. The major reason for this focus is the ease with which such analyses can be performed, as a plethora of data mining tools and techniques exist. Furthermore, the literature using such analyses is rich, providing researchers with opportunities to compare their findings. Moreover risk analysis is simply the most natural and immediately impactful application.

We also observe that there are certain areas that are sparsely filled. We attribute this emptiness

to the fact that performing research in areas such as comorbidity analysis and adverse event detection in unsupervised settings is possible but is unlikely to have significant findings. Further, conducting research in certain areas is infeasible. For example, if we are measuring the efficacy of interventions, then the very nature of the research being carried out is supervised. Also, if we are comparing two populations for specific hypothesis, the task becomes supervised in nature.

We identified a couple of reasons that research does not utilize the temporality associated with EHR data. First, the duration of EHR data available with healthcare providers rarely exceeds couple of years. Diseases such as T2DM take around 5-10 years for patients to progress from one state to a state of advanced complication. With such a limited duration of data available, this progression cannot be studied effectively. Secondly, censoring and irregular EHR data limits the application of several techniques to EHR data as such techniques often require sophisticated and rigorous study designs.

We hope that in future more advanced techniques can be developed which can model the complexity of EHR in its entirety. Robust phenotyping algorithms which can handle missing and fragmented data must be researched. The biases present in the EHR data, if any should be resolved, as they might lead to inconsistent findings. Sub-populations to be compared should have almost identical distributions for all covariates. Techniques that can handle temporal covariates and correlation between EHR data elements (e.g. laboratories test results, vitals) should be researched. We hope that our analysis provides novel insights into the way data mining research has been carried out using EHRs, thereby helping data mining to leverage its potential.

Techniques employed in mining EHRs can be borrowed from multiple sources. In this context we focus on the sources and their offerings in the context of EHR data. Epidemiology is the science that studies the patterns, causes, and effects of health and disease conditions in defined populations. It is the cornerstone of public health, and informs policy decisions and evidence-based practice by identifying risk factors for disease and targets for preventive healthcare. It encapsulates areas such as disease investigation, transmission and surveillance using case-control, cohort based and cross-sectional study designs. Data mining an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It provides the platform to identify complex biomarkers in populations which might be highly predictive for risk estimation. Survival analysis, a branch of statistics is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, etc.

Using state-of-the-art techniques from epidemiology, computer science and survival analysis can lead to discover novel methodologies. For instance, sub-population mining is one such field where all three fields can be used in an interwoven fashion. Sub-population analysis consists of techniques that compare sub-populations for notable differences. Such sub-populations can be selected based on diagnosis codes, demographic attributes, medications, and vitals. One use of such analysis is the ability to compare sub-populations for progression to advanced complications, when both sub-populations have similar medical conditions at baseline. Such comparisons can help in evaluating the effectiveness of interventions across subpopulations thereby leading to better healthcare management policies. It can also help clinicians tailor treatment to specific groups or sub-populations. EHRs, with their increased sample sizes, provide an opportunity to analyze subpopulations systematically. Epidemiological studies can be used to infer the correct study-designs, data mining for identification of sub-population cohorts, and survival analysis for modeling the outcome variables in individual subpopulations.

More recently, research has also been carried out by taking the privacy of the patients into account as due to privacy concerns and legal ramifications hospitals are often reluctant to divulge raw medical records [333–335]. Mathwe and Obradovic [333] [333**?**] worked on distributed prediction models while taking privacy of the patient into account. Divanis et al. [336, 337] discussed how private and sensitive patient data must be protected to address primary concerns. In their work they summarized various patient privacy approaches used for dissemination of patient data. They presented a survey of 45 privacy algorithms along with their advantages and disadvantages. Hoens et al. [338] developed recommendation system for identifying suitable physicians while taking the patient privacy into account. They discussed frameworks where in patients submit ratings in a protected form without revealing any information about their data.

Constructing meaningful features have also been explored in the context of healthcare informatics [349–352]. Luo et al. [350] proposed Scalable orthogonal regression to select low redundancy features. They also extended their technique by incorporating prior expertise knowledge. Sun et al. [349] proposed feature generation techniques for multi-dimensional temporal patient data, and adopt a localized supervised metric learning approach to arrive at a semantically sound similarity measure for retrieving patients represented in the multi-dimensional feature space.

Mining healthcare data is an emerging field. Healthcare informatics has a promising potential as it involves diseases such as T2DM and sepsis, for which better management practices still need to be discovered. This potential can be realized by using knowledge from diverse fields such as Epidemiology, survival analysis and data mining in an interwoven fashion. Intermixing of knowledge and techniques from varying fields has the potential for spurring development by producing more meaningful results. This can lead to the development of tailored and personalized treatments.

In this survey, we have discussed different applications for healthcare data and have attempted to provide an overview of the relevant literature for these applications. We also described the kind of data encapsulated in EHRs and the unique challenges associated with it. We then described the three major approaches used in handling EHRs namely censored data, irregular time series data and handling confounding via the pseudo outcome model. Using concepts borrowed from epidemiology we then presented the various study-designs and a comprehensive overview of the literature related to those study designs. Lastly, we presented our views on the current state of the art in healthcare informatics and envisioned what needs to be done in the future to realize the true potential associated with EHR data. We firmly believe that the unique nature of the data can contribute to the next epoch in data mining.

9 Case Study: Data Mining for Type-II Diabetes Melitus

Type 2 Diabetes Mellitus (T2DM) is a chronic condition, characterized by chronically elevated blood sugar levels. T2DM affects approximately 12% of Americans age 20 or older and is the seventh leading cause of death in the United States [353]. T2DM, unless managed effectively, leads to complications in almost every body system, including blindness, kidney disease, and various cardio-vascular complications such as peripheral vascular disease (PVD), Ischemic heart diseases (IHD), cardio vascular disease (CVD), and congestive heart failure (CHF). Effective preventive and management techniques through life style changes and therapeutic interventions exist, hence timely identification of patients at particularly high risk of developing T2DM or its complications are of paramount importance.

T2DM is part of the metabolic syndrome, a constellation of conditions related to metabolism. Beside T2DM, the metabolic syndrome contains the above complications of the diabetes, as well as a number of conditions comoribid to diabetes: high blood pressure (hypertension; HTN), high cholesterol (hyperlipidemia; HL), atherosclerosis (plaque build-up in the blood vessels) and abdominal obesity.

In what follows, we will show-case how data mining can be applied towards numerous applications in the context of T2DM and the metabolic syndrome in general. In these studies, we will describe the entire data mining process starting from raw EHR data all the way to obtaining meaningful knowledge. Specifically, we highlight some issues related to the construction of the study cohort, the synthesis of raw EHR data tables into more meaningful data elements through phenotyping, the transformation and summarization of EHR and phenotype data into a design matrix amenable to data mining through various study designs and finally we will highlight how data mining can utilize the large population samples to extract novel knowledge from the data. Most of the studies we describe in this section has actually been carried out to completion, either by us or by other researchers, but some of them are hypothetical, simply illustrating the possibilities that data mining enables.

For our discussion, we assume a typical EHR data set comprised of tables corresponding to demographics, encounters, diagnoses, laboratory results, vital signs and medication prescriptions.

Demographic Attributes This consists of patient attributes such as age, gender, race, ethnicity, socio-econmic status and tobacco consumption status. These attributes mostly remain static throughout the study period.

Encounters This contains information related to every patient visit (encounter) to the healthcare provider. Encounters are often classified as outpatient, inpatient or emergency. For every encounter, information such as encounter type, admission date, discharge date and discharge status is stored.

Diagnoses This consists of information related to newly diagnosed or existing diseases. For every diagnoses code, information such as the onset date and the date of cure (if applicable) is stored. Examples of diagnosis codes present in our dataset include codes of Type 1 and Type 2 DM, and their accompanied complications such as ischemic heart disease (IHD), cerebrovascular disease (CVD), chronic kidney disease (CKD), congestive heart failure (CHF), peripheral vascular disease (PVD), Diabetic Foot, and Ophthalmic complications.

Vitals This consists of information related to vitals, which are collected for every encounter. Information such as systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse, and body mass index (BMI). Vitals are gathered once for every outpatient visit, but might be collected frequently (every few hours or minutes) for inpatient visits, depending upon the patient medical state.

Laboratories test results For every encounter, we also store information related with various laboratories tests carried on the patient. For every entry we store information when the laboratory test result was ordered and entered into the EHR system. Examples of laboratory tests related with T2DM are hemoglobin A1c, low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), triglycerides, etc.

Prescriptions For every encounter, prescription information is also stored in the EHR. Examples of prescription are life-style modification advises and medications. For medications, information such as dosage, route, strength, prescription start date, prescription end date are usually collected.

Cohort construction.

While some clinical questions concern the entire population, most questions need to be limited to a subset of patients. For example, to determine the prevalence of T2DM, we can consider the entire population, but to understand the effect of nursing guidelines [354] we may focus on patients who were hospitalized and thus exposed to these guidelines. Accidentally including patients who were not exposed would underestimate the efficacy of the guidelines.

Cohorts are defined using *inclusion* and *exclusion* criteria governing which patients must or must not be included into the study cohort. The goal of cohort construction is to define inclusion and exclusion criteria such that the resultant cohort allows to estimate the quantities of interest without bias. In the above example, including patients without exposure could bias the estimate of the guideline efficacy.

The bias that the criteria may introduce can be obvious or subtle. When estimating the prevalence of diabetes as a ratio of diabetic patients among all patients at a provider, accidentally including patients who had died earlier introduces a negative bias (we underestimate the prevalence). On the contrary, if we were to estimate the effect of statin on mortality, we may require that patients had been taking statin for at least half a year to ensure that statin took effect. This introduces immortality bias [355], since we excluded all patients who may have died in the first half year of statin exposure and possible overestimate the beneficial effect of statin use.

In studying the effect of risk factors in diabetes, a more subtle kind of bias can aries from including patients who have a different mechanism of diabetes. For example one can reasonably argue that T2DM, formerly known as late onset diabetes, has a different mechanism in children than in adults; or investigators routinely exclude patients with gestational diabetes, a transient form of diabetes during pregnancy, for the possibility that it may have a different disease mechanism.

For these reasons, our study, we exclude children, do not include gestational diabetes and do not require a minimal set exposure time when we measure mortality.

Phenotyping. The first step in the analysis is to accurately define the clinical conditions of interest, in other words to define the disease phenotypes of interest. We illustrate this process through the example of type-II diabetes mellitus, but all other conditions should be defined analogously.

The most obvious way to identify patients with T2DM is to use diagnosis codes. There are multiple ICD-9 codes associated with T2DM depending on its severity (controled/uncontroled) and possible complications. Predefined groups of codes, such as the Clinical Classification Software [356], corresponding to diseases exist, and can be used to identify patients regardless of disease severity and complications. Identifying patients based on diagnosis codes is imperfect. A recent large multi-site study has shown that T2DM phenotype defined solely by diagnosis codes can only achieve 86.6% precision and 96.9% recall [357].

Another important goal for phenotyping is to harmonize disease definitions across time. The clinical criterion for diabetes has changed [358] thus the same diagnosis code (of T2DM) referred to a slightly different condition in the early 1990s than today. Beside the changing criterion, the laboratory tests for establishing diabetes are changing, improving. Today, the primary test for measuring blood sugar levels is hemoglobin a1c, while a mere decade earlier is was primarily fasting plasma glucose, a laboratory test with substantially higher variability. Therefore, for a longitudinal study, we have to cope with the difficulty that the same condition may have to be defined using different synonymous laboratory tests of varying accuracy.

Finally, phenotyping algorithms can help overcome challenges posed by missing data. For example, we may not have the opportunity to see the laboratory test results that established T2DM for a particular patient, either due to (left) censoring or to fragmentation, but the presence of a diagnosis code or T2DM medications can provide reliable indications of diabetes.

Phenotyping algorithms thus combine evidence from multiple data sources, diagnosis codes, laboratory results and medications, to achieve the accuracy required by the study. Phenotyping algorithms can be hand-crafted or machine learned and examples of T2DM phenotyping algorithms include [357, 359, 360].

It is also worth pointing out that even the most straightforward condition, mortality, is typically not directly available from EHR. If a patient died in the hospital, his discharge status will contain this information; however, if he died outside the hospital, his death information may not be readily available from the EHR. If mortality is of interest, researchers need to ascertain the patients' vitality status by consulting the state's population center, the state death registry or the national death registry.

Study Design
The phenotyping algorithms can be used to augment the raw EHR data with data elements that synthesize information from disparate sources and harmonize disease definition across time. The application of phenotyping algorithms to the raw EHR data (and possibly other auxiliary data) results in longitudinal data indicating whether a phenotype is confirmed, can be ruled out or cannot be established for a patient at each point in time when the patient was under observation.

For data mining algorithms to be applicable to EHR data, these tables and the phenotyping data need to be integrated and possibly summarized over time into a single design matrix. The way the design matrix is constructed is dictated by the study design, and in return, the study design constrains the applicable data mining approaches and techniques and can limit or enable certain kinds of knowledge to be extracted. For this reason, we organize the remainder of this section based on study design, presenting examples using descriptive, cross-sectional and cohort studies briefly showing how the study design drives the creation of the design matrix and how it allows the extraction of novel knowledge.

9.1 Descriptive Analysis

Descriptive studies typically represent the first forays into exploring a condition, but can also provide useful epidemiological information about diseases and thus about population health, trends in population health, thereby driving policy decisions. The Center for Disease Control and Prevention (CDC) conducts numerous descriptive analyses, annually reporting the prevalence, incidence rate and trends in diseases that represent major health care concerns, having raised attention to the growing obesity epidemic and the subsequent increase in T2DM incidence rates.

Determining prevalence and incident rates in a population of patients appears deceivingly simple, however, care must be taken with EHR data. To measure the prevalence of diabetes, we take a cross-section of the target population at a particular point in time. Prevalence is the ratio of patients who has T2DM among all patients in that target population. Phenotyping algorithms can help overcome EHR issues related to determining whether a patient has T2DM or not, but estimating the size of the target population can remain problematic. The biggest problem is selection bias. Healthy patients who require care infrequently, may not have visited the provider during the time period of the cross-section

and thus we may not know whether they are still part of the target population or not—they may have moved out of the catchment area of the provider. Whether a patient is part of the target population may be difficult to determine for the frail and the elderly, because vitality status is not necessarily available from the EHR directly.

Computing incidence rates, which are the number newly diagnosed patients during a time period divided by all patients eligible for the study during that time period, is further complicated by the need of determining whether a condition is new or pre-existent. Phenotyping algorithms can help mitigate this problem, but we can still only *estimate* the number of incident T2DM events rather than count them.

9.2 Comorbidity Analysis through Cross-Sectional Design

Comorbidity analysis is the process of exploring and analyzing relationships between frequently co-occurring diseases. For example, patients diagnosed with T2DM have often accompanied diseases such as hypertension, hyperlipidemia and impaired fasting glucose (IFG). In the aforementioned example, T2DM is an referred as the index disease and hypertension, hyperlipidemia and IFG are collectively known as co-occurring diseases. T2DM in conjunction with hypertension and hyperlipidemia are known as multiple chronic conditions (MCC). MCC's are an issue of growing significance in T2DM as they are highly prevalent and might increase disease burden and costs. Exploring and analyzing such MCC clusters will lead to development of tailored medical interventions.

The fundamental epidemiological metrics we computed above give a concise description of the health of the population and influences policy decisions, but applying data mining to it can extract deeper knowledge. In the face of an aging US population and the rapidly growing concern of multiple chronic conditions, comorbidity analysis can help describe diabetic populations in terms of comorbidities related to the metabolic syndrome, interactions among these comorbidites and estimate the prevalence and incidence rates in subpopulations defined by these chronic conditions.

The goal of this study is to identify frequently co-occurring diseases and define sub-populations based on these co-occurring diseases. Further we estimate the risk of mortality associated with each subpopulation. As this is a cross-sectional study design, we define exposure and outcome at one time point. In this study, the exposure characterized by the set of comorbidities and the outcome is defined by mortality. Using this nomenclature, we estimate the prevalence of mortality within each sub-population. We also compare analyze how the risk varies across sub-populations.

We applied frequent pattern mining to identify frequently co-occuring comorbidities and identified patient subpopulations who are diagnosed with these comorbidities (and possibly others). In each subpopulation we measure how many patients succumb to death (adjusted for age and gender) and use the Poisson test to identify subpopulations wherein the prevalence of mortality is significantly higher (or lower) than in the general population. To estimate the risk, we use Cox proportional hazards regression along with martingale residuals.

As an illustration, in the figure below we consider T2DM along with two other comorbid diseases i.e. hypertension and hyperlipidemia. We analyze the risk associated with mortality for these comorbid diseases. As observed, risk for mortality associated with hypertension and hyperlipidemia is 1.13. It indicates that the patients diagnosed with hypertension and hyperlipidemia are 13% more prone to mortality as compared to patients with no disease. Similarly we also observe how risk increases when

a patient is diagnosed with multiple diseases. The study identified a number of subpopulations with significantly elevated prevalence of diabetes. With increasing number of comorbid conditions typically, the prevalence of mortality increases, unless the combination in questions carries a particularly high risk of mortality. The increase in risk appears non-additive, suggesting interaction among the conditions under study. This is not surprising given that these conditions collectively are indicative of the patients' metabolic health.

Risk prediction for events of interest is usually performed using data mining techniques such as predictive modeling. Primarily, predictive modeling has two goals i.e. estimating the risk or identifying the underlying risk factors. For example, risk can be estimated for events such as mortality, CVD, IHD and PVD. For such estimation age, gender, race and ethnicity are the usual predictors. Data mining in EHRs also enables subpopulation mining, which helps to to build clinical decision support systems for individualized or personalized medicine. As we have already discussed that in in a cross-sectional study, we can estimate the prevalence of a disease in a population (or in well-defined subpopulations) and we can identify conditions (comorbidities) that frequently co-occur with the disease of interest. Co-occurrence is the weakest form of association; it does not even guarantee that "exposure", the development of a comorbid condition, precedes the index condition (T2DM). In a cohort study, a patient cohort is defined along with their exposures, the cohort is then followed recording outcomes of interest. This design ascertains that the exposure precedes the outcome and it also suggests that the outcome is an incident (not pre-existing) condition. Through cohort studies, we identify exposures that are *predictive* of the outcome, an application of data mining know as *biomarker discovery*, and predict the risk of the outcome (*risk estimation*). In this subsection, we would be exploring various risk estimation models such as framingham score, estimating T2DM risk in subpopulations and developing risk trajectories over time using cohort study designs.

9.2.1 Framingham Score

Let us start our discussion of cohort studies towards risk estimation with the venerable Framingham Diabetes Score [361] The Framingham Diabetes Score is a clinical tool for assessing patients' risk of developing diabetes based on a small number of risk factors: fasting blood sugar, high cholesterol, high blood pressure, medication for high blood pressure, familial history of diabetes, and obesity. For each risk factor the patient presents with, he receives a predetermined number of points. The points are tallied up and whether preventive intervention is required and the aggressiveness of the intervention is determined based on the tallied score.

In this study they estimated the 7-year risk of T2DM in middle-aged participants who had an oral glucose tolerance test at baseline. As this is a cohort study design, patients are selected based on whether they did not acquire T2DM at baseline and are usually followed for a couple of years to analyze the outcome. Patients who used oral hypoglycemic medications or insulin, or who had a baseline fasting plasma glucose level greater than 126 mg/dL or a baseline post-OGTT plasma glucose level greater than 200 mg/dL were categorized as having diabetes and thus were not included in the study. Patients were followed up from baseline for an average follow-up of 7 years. Such study designs helps in analyzing the incidence rate of T2DM.

New cases of diabetes were identified using the examination visit date as a date of diagnosis; otherwise follow-up was censored at the last follow-up (examination 6 or 7) for patients remaining nondiabetic. They used logistic regression models to predict incident diabetes and estimated the odds ratio and 95% confidence intervals to estimate relative risk. Cox proportional hazards models was also

used to account for censoring. The significant predictors identified from Cox and logistic models were similar.

They observed how parental diabetes, obesity, and metabolic syndrome traits effectively predict T2DM risk in a middle-aged white population sample. They observed how information beyond personal awareness of diabetes risk factors is important to determine risk of T2DM. They presented how parental history of diabetes and obesity remained significant predictors, along with hypertension, low levels of high-density lipoprotein cholesterol, elevated triglyceride levels, and impaired fasting glucose findings. Given the importance of identifying patients at high risk of diabetes, many risk scores have been proposed [362], but the Framingham score is the one with widest acceptance in clinical practice.

9.2.2 Diabetes Risk Prediction in Subpopulations

Clinical acceptance of the Framingham Diabetes Score is in large part due to its effectiveness (it has been validated empirically and formally [361]) and its ease of application. Its ease of application stems in large part from its approach of fitting a single model with few variables to an entire cohort, assuming homogeneity of effect across the population. In Section 9.2, we have shown through comorbidity analysis, that the comorbidities in diabetes interact. In the current study, we repeat the previous analysis using cohort study so that we can estimate diabetes risk through incident rates.

The cohort study design was applied in this study. Similarly to the cross-sectional design, a cross section was taken at a particular point in time, called the *baseline*. Demographic information (age, gender) and social history (smoking status) were determined at baseline and same comorbidities as before were ascertained retrospectively over 5 years. Patients were followed forward until 2014 and the study endpoint (outcome) was incident T2DM. Patients less than 18 years of age and patients presenting with T2DM at baseline were excluded. The latter condition ensures that all diabetes events during the follow-up period are incident (new) T2DM diagnoses.

The analysis itself mirrors that of the comorbidity analysis described earlier. Survival association rule mining [363] was applied to discover subpopulations adjusted for age, gender and follow-up time and subpopulations with increased risk of developing diabetes (i.e. incidence rate) were identified. Confounding from age and gender were handled through survival regression, which is an integral part of survival association rule mining. Unlike results from the cross-sectional design, results from this study allow us to claim that certain combinations of comorbidities are associated with higher risk of developing diabetes.

9.2.3 Quantifying the Effect of Statin

We have so far utilized data mining to identify subpopulations with significantly elevated prevalence and incidence rate of diabetes. Subpopulation mining is not limited to mining outcomes, it can also be used to discover important differences in the effects of interventions.

Recent changes in guidelines for preventing cardio-vascular mortality are expected to substantially increase the utilization of statins, a class of cholesterol lowering agent. Statins have been previously proven to reduce the risk of cardio-vascular mortality, but have been shown to increase the risk of diabetes by 9% in patients with normal blood sugar levels. Controversy surrounds the effect of statins in patients with prediabetes, a condition defined by slightly elevated sugar levels that do not reach diabetic levels [364] Most studies have found statin to have no effect on progression to overt diabetes,

some found it to be beneficial [365] and some found it to be detrimental [366].

We hypothesize that prediabetes is heterogeneous: in some subpopulation, the effect of statin is beneficial, in others it is detrimental and thus the combined effect depends on the composition of the population. In this section, we describe a study [367] that investigated the effect of statin in various subpopulations. A unique strength of this study is its importance in ascertaining subpopulations where the effect of statin is detrimental. We illustrate this process by using rigorous data mining techniques.

As this is a cohort study design, patients were selected at baseline and were followed for a couple of years to analyze the outcome. Patients were divided into treatment and control on the basis of whether they received statin or not. The groups were then followed for 5 years to estimate the incidence of T2DM. Such study designs help to examine the relationship between statin use and diabetes thereby helping to identify risk and novel protective factors.

We illustrate this process by using Association rule mining (ARM) framework in conjunction with propensity score matching techniques. Primarily, ARM was used to identify subpopulations where the effect of statins differ among subpopulations. Statistically appealing techniques such as propensity score matching was used to handle subtle biases and confounding arising due to attributes such as age and gender. Such techniques aim at eliminating the likelihood of bias and errors.

They discovered how statins substantially increase the risk of diabetes by 13% - 41% among various subpopulations. They discovered several interesting associations such as patients diagnosed with hyperlipidemia, a prescription for a non-statin anti-hyperlipidemia medication, and either obesity or treated and controlled hypertension, also receiving statins tends to lower their risk of developing diabetes. Identification of such rules are also interesting as they are easily interpretable and could be quickly incorporated into clinical practice using computer based decision support tools.

9.2.4 Trajectory Mining for Diabetes Complications

Analogous to our last study, the focus of this study is also on T2DM. Multiple studies have indicated that T2DM is often associated with several complications. Primarily, we consider seven major complications associated with diabetes: obesity (OB), ischemic heart disease (IHD), cardiovascular disease (CVD), peripheral vascular disease (PVD), cerebrovascular disease (CVD), chronic kidney disease (CKD), congestive heart failure (CHF), diabetic foot and ophthalmic conditions. These complications were identified by several research studies which dominate the literature. These complications usually stem from mismanagement of patient's health.

The aim of this case study is to analyze the risk associated with diabetes induced complications and to ascertain whether the risk changes over time. Risk can be concisely described as the probability of a subject diagnosed with T2DM progressing to a T2DM induced complication. Such analysis is also amenable for development of novel EBP (Evidence Based Guidelines) guidelines as existing EBP guidelines neither consider the patient's trajectory nor the patient's sequence of events that lead up to the patient's current conditions.

In what follows, we will illustrate how patient's risk of progressing to advanced complications depends on their present conditions. Specifically, we highlight how such analysis becomes more relevant in a heterogeneous disease such as T2DM, where in complications affect majorly all body organs. This case study provides a logical sequence from the computation of a risk score to analyzing trajectories

over time. It provided a snapshot of first foray into exploring T2DM associated complications over time.

As this is a cohort study design, patients diagnosed with T2DM at baseline (exposure) were followed for a couple of years to analyze the outcome (patient's progression to advanced complications). Patients were selected at baseline, if they satisfied the following two conditions: if they had type 1 or type 2 DM at baseline as identified by the billing transactions and two A1c results at least 6 months apart after baseline. Patients were then followed until a maximum period of 5 years or censoring or mortality (whichever occurs first) to estimate the incidence of T2DM accompanied complications. These constraints ensured that sufficient clinical information was available about the patient.

They first summarize the patient's condition pertinent to diabetes mellitus Type-2 (T2DM) into a single score using a complication index. For every complication, a Cox proportional hazards model where in demographics, laboratories test results,vitals and remaining complications are treated as the independent variables and the complication of interest as the dependent variable. Each of the individual regression models (one for each complication) provided an estimate of the coefficients, which can be interpreted as the relative risk of developing the complication of interest. These individual regression models enable the computation of Diabetes Mellitus Complication Index (DMCI) index which can be thought of as approximately 7 times the relative risk a patient faces in developing a complication. DMCI can be considered as a snapshot of patient's risk. Trajectories per complication were built by averaging the risk of patients who were diagnosed with the complication of interest. Trajectories were created using appealing statistical approaches such as spline regression and lowness estimators.

They illustrated how certain subpopulations have different risk at baseline and how certain subpopulations have substantial increase of risk in the follow up years. They also presented how different complications have varying risks of developing additional complications. For example, patients diagnosed with diabetic foot have an elevated risk of developing secondary complications. They mentioned that these patient subpopulations differ not only in their risk but also in the temporal behavior of their risk. Lastly they observed how when patients are stratified within the same subpopulation by their baseline risk, they exhibit different trajectories. Their findings lay stress on how timely analysis can help to prevent or delay the onset of accompanied complications thereby mitigating the effect of such complications on patient's health. However, the archives heel is that the research was carried out using only one dataset. Nonetheless the findings were insightful and can be validated across institutions using interoperable nature of EHR data.

10 Acknowledgements

This study is supported by National Science Foundation (NSF) grant: IIS-1344135 and by National Institute of Health (NIH) grant: LM011972. Contents of this document are the sole responsibility of the authors and do not necessarily represent official views of the NSF/NIH. This was partially supported by Grant Number 1UL1RR033183 from the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to the University of Minnesota Clinical and Translational Science Institute (CTSI).

References

[1] Hardeep Singh, Aanand Dinkar Naik, Raghuram Rao, and Laura Ann Petersen. Reducing diagnostic errors through effective communication: harnessing the power of information technology.

Journal of General Internal Medicine, 23(4):489–494, 2008.

[2] Hardeep Singh and Mark Graber. Reducing diagnostic error through medical home-based primary care reform. *Jama*, 304(4):463–464, 2010.

[3] Hardeep Singh, Mark L Graber, Stephanie M Kissam, Asta V Sorensen, Nancy F Lenfestey, Elizabeth M Tant, Kerm Henriksen, and Kenneth A LaBresh. System-related interventions to reduce diagnostic errors: a narrative review. *BMJ quality & safety*, 21(2):160–170, 2012.

[4] Erika L Abramson, Yolanda Barrón, Jill Quaresimo, and Rainu Kaushal. Electronic prescribing within an electronic health record reduces ambulatory prescribing errors. *Joint Commission Journal on Quality and Patient Safety*, 37(10):470–478, 2011.

[5] Abha Agrawal and Winfred Y Wu. Reducing medication errors and improving systems reliability using an electronic medication reconciliation system. *Joint Commission Journal on Quality and Patient Safety*, 35(2):106–114, 2009.

[6] David Dorr, Laura M Bonner, Amy N Cohen, Rebecca S Shoai, Ruth Perrin, Edmund Chaney, and Alexander S Young. Informatics systems to promote improved care for chronic illness: a literature review. *Journal of the American Medical Informatics Association*, 14(2):156–163, 2007.

[7] James B Meigs, Enrico Cagliero, Anil Dubey, Patricia Murphy-Sheehy, Catharyn Gildesgame, Henry Chueh, Michael J Barry, Daniel E Singer, and David M Nathan. A controlled trial of web-based diabetes disease management the mgh diabetes primary care improvement project. *Diabetes Care*, 26(3):750–757, 2003.

[8] Stephen M Shortell, Robin Gillies, Juned Siddique, Lawrence P Casalino, Diane Rittenhouse, James C Robinson, and Rodney K McCurdy. Improving chronic illness care: a longitudinal cohort analysis of large physician organizations. *Medical care*, 47(9):932–939, 2009.

[9] Bruce L Rollman, Barbara H Hanusa, Henry J Lowe, Trae Gilbert, Wishwa N Kapoor, and Herbert C Schulberg. A randomized trial using computerized decision support to improve treatment of major depression in primary care. *Journal of General Internal Medicine*, 17(7):493–503, 2002.

[10] Jesse C Crosson, Pamela A Ohman-Strickland, Karissa A Hahn, Barbara DiCicco-Bloom, Eric Shaw, A John Orzano, and Benjamin F Crabtree. Electronic medical records and diabetes quality of care: results from a sample of family medicine practices. *The Annals of Family Medicine*, 5(3):209–215, 2007.

[11] Teri A Manolio. Collaborative genome-wide association studies of diverse diseases: programs of the nhgri's office of population genomics. 2009.

[12] Alexander G Fiks, Kenya F Hunter, A Russell Localio, Robert W Grundmeier, Tyra Bryant-Stephens, Anthony A Luberti, Louis M Bell, and Evaline A Alessandrini. Impact of electronic health record-based alerts on influenza vaccination for children with asthma. *Pediatrics*, 124(1):159–169, 2009.

[13] Angela M Davis, Matthew Cannon, Adrienne Z Ables, and Heather Bendyk. Using the electronic medical record to improve asthma severity documentation and treatment among family medicine residents. *Family medicine*, 42(5):334, 2010.

[14] Per-Olof Ehrs and Kjeu Larsson. Treatment improves quality of life in patients with poor perception of asthma. *Primary Care Respiratory Journal*, 13(1):42–47, 2004.

[15] Louis M Bell, Robert Grundmeier, Russell Localio, Joseph Zorc, Alexander G Fiks, Xuemei Zhang, Tyra Bryant Stephens, Marguerite Swietlik, and James P Guevara. Electronic health record–based decision support to improve asthma care: a cluster-randomized trial. *Pediatrics*, 125(4):e770–e777, 2010.

[16] Valerie Weber, Frederick Bloom, Steve Pierdon, and Craig Wood. Employing the electronic health record to improve diabetes care: a multifaceted intervention in an integrated delivery system. *Journal of general internal medicine*, 23(4):379–382, 2008.

[17] Patrick J OConnor, JoAnn M Sperl-Hillen, William A Rush, Paul E Johnson, Gerald H Amundson, Stephen E Asche, Heidi L Ekstrom, and Todd P Gilmer. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. *The Annals of Family Medicine*, 9(1):12–21, 2011.

[18] Joel Kupersmith, Joseph Francis, Eve Kerr, Sarah Krein, Leonard Pogach, Robert M Kolodner, and Jonathan B Perlin. Advancing evidence-based care for diabetes: lessons from the veterans health administration. *Health Affairs*, 26(2):w156–w168, 2007.

[19] Randall D Cebul, Thomas E Love, Anil K Jain, and Christopher J Hebert. Electronic health records and quality of diabetes care. *New England Journal of Medicine*, 365(9):825–833, 2011.

[20] Linda T Bilheimer and Richard J Klein. Data and measurement issues in the analysis of health disparities. *Health services research*, 45(5p2):1489–1507, 2010.

[21] Douglas W Roblin, Thomas K Houston, Jeroan J Allison, Peter J Joski, and Edmund R Becker. Disparities in use of a personal health record in a managed care organization. *Journal of the American Medical Informatics Association*, 16(5):683–689, 2009.

[22] Catherine Lejeune, Franco Sassi, Libby Ellis, Sara Godward, Vivian Mak, Matthew Day, and Bernard Rachet. Socio-economic disparities in access to treatment and their impact on colorectal cancer survival. *International journal of epidemiology*, 39(3):710–717, 2010.

[23] John NS Matthews. *Introduction to randomized controlled clinical trials*. CRC Press, 2006.

[24] Kenneth F Schulz and David A Grimes. Sample size calculations in randomised trials: mandatory and mystical. *Lancet*, 365(9467):1348–53.

[25] An-Wen Chan, Asbjø rn Hróbjartsson, Karsten J Jø rgensen, Peter C Gø tzsche, and Douglas G Altman. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ (Clinical research ed.)*, 337:a2299, January 2008.

[26] E Passamani, K B Davis, M J Gillespie, and T Killip. A randomized trial of coronary artery bypass surgery. Survival of patients with a low ejection fraction. *The New England journal of medicine*, 312(26):1665–71, June 1985.

[27] Harriette G C Van Spall, Andrew Toren, Alex Kiss, and Robert A Fowler. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA : the journal of the American Medical Association*, 297(11):1233–40, March 2007.

[28] M Olschewski, M Schumacher, and K B Davis. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled clinical trials*, 13(3):226–39, June 1992.

[29] W S Moore, B Young, W H Baker, J T Robertson, J F Toole, C L Vescera, and V J Howard. Surgical results: a justification of the surgeon selection process for the ACAS trial. The ACAS Investigators. *Journal of vascular surgery*, 23(2):323–8, February 1996.

[30] J H Gurwitz, N F Col, and J Avorn. The exclusion of the elderly and women from clinical trials in acute myocardial infarction. *JAMA : the journal of the American Medical Association*, 268(11):1417–22, September 1992.

[31] Peter M Rothwell. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*, 365(9453):82–93.

[32] R L Sacco, D E Kargman, Q Gu, and M C Zamanillo. Race-ethnicity and determinants of intracranial atherosclerotic cerebral infarction. The Northern Manhattan Stroke Study. *Stroke; a journal of cerebral circulation*, 26(1):14–20, January 1995.

[33] Martin Fortin, Jonathan Dionne, Geneviève Pinho, Julie Gignac, José Almirall, and Lise Lapointe. Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Annals of family medicine*, 4(2):104–108.

[34] Ajay K Kakkar, Benjamin Brenner, Ola E Dahl, Bengt I Eriksson, Patrick Mouret, Jim Muntz, Andrea G Soglian, Akos F Pap, Frank Misselwitz, and Sylvia Haas. Extended duration rivaroxaban versus short-term enoxaparin for the prevention of venous thromboembolism after total hip arthroplasty: a double-blind, randomised controlled trial. *Lancet*, 372(9632):31–9, July 2008.

[35] Keith Feldman and Nitesh V Chawla. Scaling personalized healthcare with big data. In *2nd International Conference on Big Data and Analytics in Healthcare, Singapore*, 2014.

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

[37] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

[38] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[39] Pascal Coorevits, M Sundgren, GO Klein, A Bahr, B Claerhout, C Daniel, M Dugas, D Dupont, A Schmidt, P Singleton, et al. Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6):547–560, 2013.

[40] Fabricio F Costa. Big data in biomedicine. *Drug discovery today*, 19(4):433–440, 2014.

[41] MK Ross, Wei Wei, and L Ohno-Machado. big data and the electronic health record. *Yearbook of medical informatics*, 9(1):97, 2014.

[42] Chandan K Reddy and Charu C Aggarwal. *HealtHcare Data analytics*, volume 36. CRC Press, 2015.

[43] Francisco S Roque, Peter B Jensen, Henriette Schmock, Marlene Dalgaard, Massimo Andreatta, Thomas Hansen, Karen Sø eby, Sø ren Bredkjæ r, Anders Juul, Thomas Werge, Lars J Jensen, and Sø ren Brunak. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS computational biology*, 7(8):e1002141, August 2011.

[44] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–63, January 2014.

[45] Adam Wright, Elizabeth S Chen, and Francine L Maloney. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics*, 43(6):891–901, December 2010.

[46] Hui Cao, Marianthi Markatou, Genevieve B Melton, Michael F Chiang, and George Hripcsak. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 106–110, January 2005.

[47] Antony B Holmes, Alexander Hawson, Feng Liu, Carol Friedman, Hossein Khiabanian, and Raul Rabadan. Discovering disease associations by integrating electronic clinical data and medical literature. *PloS one*, 6(6):e21132, January 2011.

[48] A Mi Shin, In Hee Lee, Gyeong Ho Lee, Hee Joon Park, Hyung Seop Park, Kyung Il Yoon, Jung Jeung Lee, and Yoon Nyun Kim. Diagnostic analysis of patients with essential hypertension using association rule mining. *Healthcare informatics research*, 16(2):77–81, June 2010.

[49] Dipanwita Dasgupta and Nitesh V Chawla. Disease and medication networks: An insight into disease-drug interactions.

[50] Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5, 2014.

[51] Joan M Teno, Sherry Weitzen, Mary L Fennell, and Vincent Mor. Dying trajectory in the last year of life: does cancer trajectory fit other diseases? *Journal of palliative medicine*, 4(4):457–464, 2001.

[52] Fliss EM Murtagh, Emma Murphy, and Neil S Sheerin. Illness trajectories: an important concept in the management of kidney failure. *Nephrology Dialysis Transplantation*, 23(12):3746–3748, 2008.

[53] Jeff Friedlin, Marc Overhage, Mohammed A Al-Haddad, Joshua A Waters, J Juan R Aguilar-Saavedra, Joe Kesterson, and Max Schmidt. Comparing methods for identifying pancreatic cancer patients using electronic data sources. In *AMIA Annual Symposium Proceedings*, volume 2010, page 237. American Medical Informatics Association, 2010.

[54] Katherine P Liao, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-treitler, Soumya Raychaudhuri, Peter Szolovits, Susanne Churchill, Shawn Murphy, Isaac Kohane, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*, 62(8):1120–1127, 2010.

[55] Yoni Halpern, Youngduck Choi, Steven Horng, and David Sontag. Using anchors to estimate clinical state without labeled data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 606. American Medical Informatics Association, 2014.

[56] Stéphane M Meystre, Vikrant G Deshmukh, and Joyce Mitchell. A clinical use case to evaluate the i2b2 hive: predicting asthma exacerbations. In *AMIA Annual Symposium Proceedings*, volume 2009, page 442. American Medical Informatics Association, 2009.

[57] Sasikiran Kandula, Qing Zeng-Treitler, Lingji Chen, William L Salomon, and Bruce E Bray. A bootstrapping algorithm to improve cohort identification using structured data. *Journal of biomedical informatics*, 44:S63–S68, 2011.

[58] Luke V Rasmussen, Will K Thompson, Jennifer A Pacheco, Abel N Kho, David S Carrell, Jyotishman Pathak, Peggy L Peissig, Gerard Tromp, Joshua C Denny, and Justin B Starren. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *Journal of biomedical informatics*, 51:280–286, 2014.

[59] William P Castelli, Joseph T Doyle, Tavia Gordon, CURTIS G Hames, MARTHANA C Hjortland, STEPHEN B Hulley, A Kagan, and WILLIAM J Zukel. Hdl cholesterol and other lipids in coronary heart disease. the cooperative lipoprotein phenotyping study. *Circulation*, 55(5):767–772, 1977.

[60] Katherine M Newton, Peggy L Peissig, Abel Ngo Kho, Suzette J Bielinski, Richard L Berg, Vidhu Choudhary, Melissa Basford, Christopher G Chute, Iftikhar J Kullo, Rongling Li, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the emerge network. *Journal of the American Medical Informatics Association*, 20(e1):e147–e154, 2013.

[61] Casey Lynnette Overby, Jyotishman Pathak, Omri Gottesman, Krystl Haerian, Adler Perotte, Sean Murphy, Kevin Bruce, Stephanie Johnson, Jayant Talwalkar, Yufeng Shen, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. *Journal of the American Medical Informatics Association*, 20(e2):e243–e252, 2013.

[62] Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(e2):e206–e211, 2013.

[63] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald MA Henry, and Coen DA Stehouwer. The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451, 2014.

[64] Mary Regina Boland, George Hripcsak, Yufeng Shen, Wendy K Chung, and Chunhua Weng. Defining a comprehensive verotype using electronic health records for personalized medicine. *Journal of the American Medical Informatics Association*, 20(e2):e232–e238, 2013.

[65] David Gotz, Fei Wang, and Adam Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of biomedical informatics*, 48:148–159, 2014.

[66] Xiang Wang, Fei Wang, Jun Wang, Buyue Qian, and Jianying Hu. Exploring patient risk groups with incomplete knowledge. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1223–1228. IEEE, 2013.

[67] Peggy L Peissig, Luke V Rasmussen, Richard L Berg, James G Linneman, Catherine A McCarty, Carol Waudby, Lin Chen, Joshua C Denny, Russell A Wilke, Jyotishman Pathak, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *Journal of the American Medical Informatics Association*, 19(2):225–234, 2012.

[68] Jyotishman Pathak, Richard C Kiefer, Suzette J Bielinski, and Christopher G Chute. Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank. *J. Biomedical Semantics*, 3:10, 2012.

[69] Joyce C Ho, Joydeep Ghosh, Steve R Steinhubl, Walter F Stewart, Joshua C Denny, Bradley A Malin, and Jimeng Sun. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics*, 52:199–211, 2014.

[70] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124. ACM, 2014.

[71] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Extracting phenotypes from patient claim records using nonnegative tensor factorization. In *Brain Informatics and Health*, pages 142–151. Springer, 2014.

[72] Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[73] Jianying Hu, Fei Wang, Jimeng Sun, Robert Sorrentino, and Shahram Ebadollahi. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. In *AMIA Annual Symposium Proceedings*, volume 2012, page 360. American Medical Informatics Association, 2012.

[74] Kenney Ng, Amol Ghoting, Steven R Steinhubl, Walter F Stewart, Bradley Malin, and Jimeng Sun. Paramo: A parallel predictive modeling platform for healthcare analytic research using electronic health records. *Journal of biomedical informatics*, 48:160–170, 2014.

[75] Truyen Tran, Dinh Phung, Wei Luo, and Svetha Venkatesh. Stabilized sparse ordinal regression for medical risk stratification. *Knowledge and Information Systems*, 43(3):555–582, 2014.

[76] Theodoros Rekatsinas, Saurav Ghosh, Sumiko R Mekaru, Elaine O Nsoesie, John S Brownstein, Lise Getoor, and Naren Ramakrishnan. Sourceseer: Forecasting rare disease outbreaks using multiple data sources. *Timeline*, 7:8.

[77] Philip Greenland, Laurie LaBree, Stanley P Azen, Terence M Doherty, and Robert C Detrano. Coronary artery calcium score combined with framingham score for risk prediction in asymptomatic individuals. *Jama*, 291(2):210–215, 2004.

[78] William A Knaus, Douglas P Wagner, Elizabeth A Draper, Jack E Zimmerman, Marilyn Bergner, Paulo G Bastos, Carl A Sirio, Donald J Murphy, Ted Lotring, and Anne Damiano. The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. *Chest Journal*, 100(6):1619–1636, 1991.

[79] Chandrima Sarkar, Sarah Cooley, and Jaideep Srivastava. Improved feature selection for hematopoietic cell transplantation outcome prediction using rank aggregation. In *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*, pages 221–226. IEEE, 2012.

[80] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. An interpretable stroke prediction model using rules and bayesian analysis. 2013.

[81] Shahram Ebadollahi, Jimeng Sun, David Gotz, Jianying Hu, Daby Sow, and Chalapathy Neti. Predicting patients trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics. In *AMIA annual symposium proceedings*, volume 2010, page 192. American Medical Informatics Association, 2010.

[82] Keith Feldman and Nitesh V Chawla. Admission duration model for infant treatment (admit). In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 583–587. IEEE, 2014.

[83] Che Ngufor, Sudhindra Upadhyaya, Dennis Murphree, Nageswar Madde, Daryl Kor, and Jyotishman Pathak. A heterogeneous multi-task learning for predicting rbc transfusion and perioperative outcomes. In *Artificial Intelligence in Medicine*, pages 287–297. Springer, 2015.

[84] Roy J Byrd, Steven R Steinhubl, Jimeng Sun, Shahram Ebadollahi, and Walter F Stewart. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International journal of medical informatics*, 83(12):983–992, 2014.

[85] Iman Kamkar, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. Stable feature selection for clinical prediction: Exploiting icd tree structure using tree-lasso. *Journal of biomedical informatics*, 53:277–290, 2015.

[86] Truyen Tran, Wei Luo, Dinh Phung, Sunil Gupta, Santu Rana, Richard L Kennedy, Ann Larkins, and Svetha Venkatesh. A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC bioinformatics*, 15(1):425, 2014.

[87] Geetika T Lakshmanan, Szabolcs Rozsnyai, and Fei Wang. Investigating clinical care pathways correlated with outcomes. In *Business process management*, pages 323–338. Springer, 2013.

[88] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

[89] Rajakrishnan Vijayakrishnan, Steven R Steinhubl, Kenney Ng, Jimeng Sun, Roy J Byrd, Zahra Daar, Brent A Williams, Shahram Ebadollahi, Walter F Stewart, et al. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *Journal of cardiac failure*, 20(7):459–464, 2014.

[90] Pratibha Vellanki, Thi Duong, Svetha Venkatesh, and Dinh Phung. Nonparametric discovery of learning patterns and autism subgroups from therapeutic data. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1828–1833. IEEE, 2014.

[91] John R Schrom, Pedro J Caraballo, M Regina Castro, and György J Simon. Quantifying the effect of statin use in pre-diabetic phenotypes discovered through association rule mining. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2013:1249–1257, January 2013.

[92] Gyorgy J Simon, John Schrom, M Regina Castro, Peter W Li, and Pedro J Caraballo. Survival association rule mining towards type 2 diabetes risk assessment. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2013:1293–302, January 2013.

[93] Rave Harpaz, Santiago Vilar, William Dumouchel, Hojjat Salmasian, Krystl Haerian, Nigam H Shah, Herbert S Chase, and Carol Friedman. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association : JAMIA*, 20(3):413–9, May 2013.

[94] George Hripcsak, David J Albers, and Adler Perotte. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association*, page ocu051, 2015.

[95] Yubin Park and Joydeb Ghosh. Ensembles of ({*alpha*})-trees for imbalanced classification problems. *Knowledge and Data Engineering, IEEE Transactions on*, 26(1):131–143, 2014.

[96] Lisiane Pranjul Yadav, Sanjoy Andrew, Bonnie Katherine, Vipin Connie, and Gyorgy Simon Michael. Modelling trajectories for diabetes complications. *SDM*, 2015.

[97] Di Zhao and Chunhua Weng. Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction. *Journal of biomedical informatics*, 44(5):859–868, 2011.

[98] Francisco Javier Algar, Antonio Alvarez, Angel Salvatierra, Carlos Baamonde, José Luis Aranda, and Francisco Javier López-Pujol. Predicting pulmonary complications after pneumonectomy for lung cancer. *European journal of cardio-thoracic surgery*, 23(2):201–208, 2003.

[99] Allan S Detsky, Howard B Abrams, John R McLaughlin, Daniel J Drucker, Zion Sasson, Nancy Johnston, J Gerald Scott, Nicholas Forbath, and Joseph R Hilliard. Predicting cardiac complications in patients undergoing non-cardiac surgery. *Journal of general internal medicine*, 1(4):211–219, 1986.

[100] Jonathan F Finks, Kerry L Kole, Panduranga R Yenumula, Wayne J English, Kevin R Krause, Arthur M Carlin, Jeffrey A Genaw, Mousumi Banerjee, John D Birkmeyer, Nancy J Birkmeyer, et al. Predicting risk for serious complications with bariatric surgery: results from the michigan bariatric surgery collaborative. *Annals of surgery*, 254(4):633–640, 2011.

[101] DC McCrory, LB Goldstein, GP Samsa, EZ Oddone, PB Landsman, WS Moore, and DB Matchar. Predicting complications of carotid endarterectomy. *Stroke*, 24(9):1285–1291, 1993.

[102] Alastair D Hay, Tom Fahey, Tim J Peters, and Andrew Wilson. Predicting complications from acute cough in pre-school children in primary care: a prospective cohort study. *British journal of general practice*, 54(498):9–14, 2004.

[103] Colleen M McCarthy, Babak J Mehrara, Elyn Riedel, Kristen Davidge, Akili Hinson, Joseph J Disa, Peter G Cordeiro, and Andrea L Pusic. Predicting complications following expander/implant breast reconstruction: an outcomes analysis based on preoperative clinical risk. *Plastic and reconstructive surgery*, 121(6):1886–1892, 2008.

[104] SK Epstein, LJ Faling, BD Daly, and BR Celli. Predicting complications after pulmonary resection. preoperative exercise testing vs a multifactorial cardiopulmonary risk index. *CHEST Journal*, 104(3):694–700, 1993.

[105] Mark K Ferguson and Amy E Durkin. A comparison of three scoring systems for predicting complications after major lung resection. *European journal of cardio-thoracic surgery*, 23(1):35–42, 2003.

[106] Mark K Ferguson and Amy E Durkin. Preoperative prediction of the risk of pulmonary complications after esophagectomy for cancer. *The Journal of thoracic and cardiovascular surgery*, 123(4):661–669, 2002.

[107] MK Ferguson, L Little, L Rizzo, KJ Popovich, GF Glonek, A Leff, D Manjoney, and AG Little. Diffusing capacity predicts morbidity and mortality after pulmonary resection. *The Journal of thoracic and cardiovascular surgery*, 96(6):894–900, 1988.

[108] Nelson F SooHoo, Jay R Lieberman, Clifford Y Ko, and David S Zingmond. Factors predicting complication rates following total knee replacement. *The Journal of Bone & Joint Surgery*, 88(3):480–485, 2006.

[109] Robert J Benz, Zaki G Ibrahim, Pouya Afshar, and Steven R Garfin. Predicting complications in elderly patients undergoing lumbar decompression. *Clinical orthopaedics and related research*, 384:116–121, 2001.

[110] Murat Y Ozkalkanli, Dila Tuna Ozkalkanli, Kaan Katircioglu, and Serdar Savaci. Comparison of tools for nutrition assessment and screening for predicting the development of complications in orthopedic surgery. *Nutrition in Clinical Practice*, 24(2):274–280, 2009.

[111] Anthony M Propst, Rebecca F Liberman, Bernard L Harlow, and Elizabeth S Ginsburg. Complications of hysteroscopic surgery: predicting patients at risk. *Obstetrics & Gynecology*, 96(4):517–520, 2000.

[112] Jean Klastersky, Marianne Paesmans, Aspasia Georgala, Frédérique Muanza, Barbara Plehiers, Laurent Dubreucq, Yassine Lalami, Michel Aoun, and Martine Barette. Outpatient oral antibiotics for febrile neutropenic cancer patients using a score predictive for complications. *Journal of clinical oncology*, 24(25):4129–4134, 2006.

[113] Zitao Liu and Milos Hauskrecht. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial intelligence in medicine*, 2014.

[114] Zitao Liu and Milos Hauskrecht. Clinical time series prediction with a hierarchical dynamical system. In *Artificial Intelligence in Medicine*, pages 227–237. Springer, 2013.

[115] Zitao Liu, Lei Wu, and Milos Hauskrecht. Modeling clinical time series using gaussian process sequences. In *SIAM international conference on data mining*, pages 623–631. Citeseer, 2013.

[116] Zitao Liu and Milos Hauskrecht. Sparse linear dynamical system with its application in multivariate clinical time series. *arXiv preprint arXiv:1311.7071*, 2013.

[117] Maryam Panahiazar, Vahid Taslimitehrani, Naveen L Pereira, and Jyotishman Pathak. Using ehrs for heart failure therapy recommendation using multidimensional patient similarity analytics. *Studies in health technology and informatics*, 210:369–373, 2014.

[118] Suzann K Campbell. Quantifying the effects of interventions for movement disorders resulting from cerebral palsy. *Journal of child neurology*, 11(1 suppl):S61–S70, 1996.

[119] Judith J Prochaska, Wayne F Velicer, Claudio R Nigg, and James O Prochaska. Methods of quantifying change in multiple risk factor interventions. *Preventive medicine*, 46(3):260–265, 2008.

[120] Carine Ronsmans and Oona Campbell. Quantifying the fall in mortality associated with interventions related to hypertensive diseases of pregnancy. *BMC Public Health*, 11(Suppl 3):S8, 2011.

[121] Malcolm R Law, Nicholas J Wald, AR Rudnicka, et al. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *Bmj*, 326(7404):1423, 2003.

[122] Marilyn J Field, Kathleen N Lohr, et al. *Guidelines for Clinical Practice:: From Development to Use*. National Academies Press, 1992.

[123] Sistine A Barretto, Jim Warren, Andrew Goodchild, Linda Bird, Sam Heard, and Markus Stumptner. Linking guidelines to electronic health record design for improved chronic disease management. In *AMIA Annual Symposium Proceedings*, volume 2003, page 66. American Medical Informatics Association, 2003.

[124] Yueyi I Liu and Daniel L Rubin. The role of informatics in health care reform. *Academic radiology*, 19(9):1094–9, September 2012.

[125] David Eibling, Marvin Fried, Andrew Blitzer, and Gregory Postma. Commentary on the role of expert opinion in developing evidence-based guidelines. *The Laryngoscope*, 124(2):355–7, February 2014.

[126] Pooja Agrawal and Joshua M Kosowsky. Clinical practice guidelines in the emergency department. *Emergency medicine clinics of North America*, 27(4):555–567, 2009.

[127] Linan Zeng, Lingli Zhang, Zhiqiang Hu, Emily A Ehle, Yuan Chen, Lili Liu, and Min Chen. Systematic review of evidence-based guidelines on medication therapy for upper respiratory tract infection in children with agree instrument. *PloS one*, 9(2):e87711, 2014.

[128] Annemie Heselmans, Stijn Van de Velde, Dirk Ramaekers, Robert Vander Stichele, and Bert Aertgeerts. Feasibility and impact of an evidence-based electronic decision support system for diabetes care in family medicine: protocol for a cluster randomized controlled trial. *Implementation science : IS*, 8(1):83, January 2013.

[129] Avinash S Bidra. Evidence-based prosthodontics: fundamental considerations, limitations, and guidelines. *Dental clinics of North America*, 58(1):1–17, January 2014.

[130] Maritta Kinnunen-Amoroso. How occupational health care professionals experience evidence-based guidelines in Finland: a qualitative study. *Journal of evaluation in clinical practice*, 19(4):612–6, August 2013.

[131] S Y Goh, S B Ang, Y M Bee, Y T Chen, D S Gardner, E T Ho, K Adaikan, Y C Lee, C H Lee, F S Lim, H B Lim, S C Lim, J Seow, A W Soh, C F Sum, E S Tai, A C Thai, T Y Wong, and F Yap. Ministry of health clinical practice guidelines: diabetes mellitus. *Singapore medical journal*, 55(6):334–47, June 2014.

[132] Rimma Pivovarov, David J Albers, George Hripcsak, Jorge L Sepulveda, and Noémie Elhadad. Temporal trends of hemoglobin a1c testing. *Journal of the American Medical Informatics Association*, 21(6):1038–1044, 2014.

[133] Qi Li, Kristin Melton, Todd Lingren, Eric S Kirkendall, Eric Hall, Haijun Zhai, Yizhao Ni, Megan Kaiser, Laura Stoutenborough, and Imre Solti. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. *Journal of the American Medical Informatics Association : JAMIA*, 21(5):776–84, January 2014.

[134] Santiago Vilar, Rave Harpaz, Lourdes Santana, Eugenio Uriarte, and Carol Friedman. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis. *PloS one*, 7(7):e41471, January 2012.

[135] Richard H Epstein, Paul St Jacques, Michael Stockin, Brian Rothman, Jesse M Ehrenfeld, and Joshua C Denny. Automated identification of drug and food allergies entered using non-standard terminology. *Journal of the American Medical Informatics Association : JAMIA*, 20(5):962–8.

[136] Srinivasan V Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H Shah. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association : JAMIA*, 21(2):353–62.

[137] K Haerian, D Varn, S Vaidya, L Ena, H S Chase, and C Friedman. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical pharmacology and therapeutics*, 92(2):228–34, August 2012.

[138] Ryen W White, Nicholas P Tatonetti, Nigam H Shah, Russ B Altman, and Eric Horvitz. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association : JAMIA*, 20(3):404–8, May 2013.

[139] Preciosa M Coloma, Paul Avillach, Francesco Salvo, Martijn J Schuemie, Carmen Ferrajolo, Antoine Pariente, Annie Fourrier-Réglat, Mariam Molokhia, Vaishali Patadia, Johan van der Lei, Miriam Sturkenboom, and Gianluca Trifirò. A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases. *Drug safety : an international journal of medical toxicology and drug experience*, 36(1):13–23, January 2013.

[140] William V Bobo, Jyotishman Pathak, Hilal Maradit Kremers, Barbara P Yawn, Scott M Brue, Cynthia J Stoppel, Paul E Croarkin, Jennifer St Sauver, Mark A Frye, and Walter A Rocca. An electronic health record driven algorithm to identify incident antidepressant medication users. *Journal of the American Medical Informatics Association*, 21(5):785–791, 2014.

[141] Jyotishman Pathak, Richard C Kiefer, and Christopher G Chute. Mining anti-coagulant drug-drug interactions from electronic health records using linked data. In *Data Integration in the Life Sciences*, pages 128–140. Springer, 2013.

[142] Feichen Shen, Dingcheng Li, Hongfang Liu, Yugyung Lee, Jyotishman Pathak, Christopher G Chute, and Cui Tao. Using semantic web technologies for quality measure phenotyping algorithm representation and automatic execution on ehr data. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014.

[143] Aarti Sathyanarayana, Jyotishman Pathak, Rozalina McCoy, Santiago Romero-Brufau, Maryam Panaziahar, and Jaideep Srivastava. Clinical decision making: A framework for predicting rx response. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 1185–1188. IEEE, 2014.

[144] Senthil K Nachimuthu, Anthony Wong, and Peter J Haug. Modeling Glucose Homeostasis and Insulin Dosing in an Intensive Care Unit using Dynamic Bayesian Networks. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:532–6, January 2010.

[145] C. Rose, C. Smaili, and F. Charpillet. A dynamic Bayesian network for handling uncertainty in a decision support system adapted to the monitoring of patients treated by hemodialysis. In *17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)*, pages 5 pp.–598. IEEE, 2005.

[146] Melinda M Davis, Jillian M Currey, Sonya Howk, Molly R DeSordi, Linda Boise, Lyle J Fagnan, and Nancy Vuckovic. A qualitative study of rural primary care clinician views on remote monitoring technologies. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association*, 30(1):69–78, January 2014.

[147] E T van der Velde, H Foeken, T A Witteman, L van Erven, and M J Schalij. Integration of data from remote monitoring systems and programmers into the hospital electronic health record

system based on international standards. *Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation*, 20(2):66–70, February 2012.

[148] Wajahat Ali Khan, Maqbool Hussain, Muhammad Afzal, Muhammad Bilal Amin, and Sungyoung Lee. Healthcare standards based sensory data exchange for Home Healthcare Monitoring System. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2012:1274–7, January 2012.

[149] O Farri, A Rahman, KA Monsen, R Zhang, SV Pakhomov, DS Pieczkiewicz, SM Speedie, GB Melton, et al. Impact of a prototype visualization tool for new information in ehr clinical documents. *Appl Clin Inform*, 3(4):404–418, 2012.

[150] Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307, 2014.

[151] Yan Wang, Serguei Pakhomov, Nora E Burkart, James O Ryan, and Genevieve B Melton. A study of actions in operative notes. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1431. American Medical Informatics Association, 2012.

[152] Lawrence L Weed. Special article: Medical records that guide and teach. *New England Journal of Medicine*, 278(12):593–600, 1968.

[153] Halina Frydman. Nonparametric estimation of a markov illness-deathprocess from interval-censored observations, with application to diabetes survival data. *Biometrika*, 82(4):773–789, 1995.

[154] Pierre Joly, Daniel Commenges, Catherine Helmer, and Luc Letenneur. A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, 3(3):433–443, 2002.

[155] Per Kragh Andersen. Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in medicine*, 7(6):661–670, 1988.

[156] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[157] Lee-Jen Wei, Danyu Y Lin, and L Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, 84(408):1065–1073, 1989.

[158] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[159] David Moher, Alison Jones, Deborah J Cook, Alejandro R Jadad, Michael Moher, Peter Tugwell, Terry P Klassen, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *The Lancet*, 352(9128):609–613, 1998.

[160] Kenneth F Schulz, Iain Chalmers, Richard J Hayes, and Douglas G Altman. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Jama*, 273(5):408–412, 1995.

[161] Stephen B Gruber. Clinical epidemiology: The architecture of clinical research. *The Yale journal of biology and medicine*, 59(1):77, 1986.

[162] Ronald T Burkman. Association between intrauterine device and pelvic inflammatory disease. *Obstetrics & Gynecology*, 57(3):269–276, 1981.

[163] BEC Nordin. Oestrogen treatment and endometrial carcinoma. *British medical journal*, 2(6084):454, 1977.

[164] Carl C Seltzer, Raymond Bosse, and Arthur J Garvey. Mail survey response by smoking status. *American journal of epidemiology*, 100(6):453–457, 1974.

[165] Sholom Wacholder, Debra T Silverman, Joseph K McLaughlin, and Jack S Mandel. Selection of controls in case-control studies: Ii. types of controls. *American journal of epidemiology*, 135(9):1029–1041, 1992.

[166] Leonard J Vernick, Susan L Vernick, and Lewis H Kuller. Selection of neighborhood controls: logistics and fieldwork. *Journal of chronic diseases*, 37(3):177–182, 1984.

[167] E RYU JACQUELINE, CORLEEN J THOMPSON, and JOHN R CROUSE. Selection of neighborhood controls for a study of coronary artery disease. *American journal of epidemiology*, 129(2):407–414, 1989.

[168] Joseph Herbert Abramson and ZH Abramson. *Making sense of data: A self-instruction manual on the interpretation of epidemiological data*. Oxford university press, 2001.

[169] Jeffrey J Walline. Designing clinical research: an epidemiologic approach, 2001.

[170] Howard W Ory. Association between oral contraceptives and myocardial infarction: a review. *Jama*, 237(24):2619–2622, 1977.

[171] Pamela J Schwingl, Howard W Ory, and Cynthia M Visness. Estimates of the risk of cardiovascular death attributable to low-dose oral contraceptives in the united states. *American journal of obstetrics and gynecology*, 180(1):241–249, 1999.

[172] Anrudh K Jain. Cigarette smoking, use of oral contraceptives, and myocardial infarction. *American journal of obstetrics and gynecology*, 126(3):301–307, 1976.

[173] Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies. *J natl cancer inst*, 22(4):719–748, 1959.

[174] Thomas Allen Lang and Michelle Secic. *How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers*. ACP Press, 2006.

[175] David R Cox. Regression models and life-tables. In *Breakthroughs in Statistics*, pages 527–541. Springer, 1992.

[176] Charanjit S Rihal, Stephen C Textor, Diane E Grill, Peter B Berger, Henry H Ting, Patricia J Best, Mandeep Singh, Malcolm R Bell, Gregory W Barsness, Verghese Mathew, et al. Incidence and prognostic importance of acute renal failure after percutaneous coronary intervention. *Circulation*, 105(19):2259–2264, 2002.

[177] John A Dormandy, Bernard Charbonnel, David JA Eckland, Erland Erdmann, Massimo Massi-Benedetti, Ian K Moules, Allan M Skene, Meng H Tan, Pierre J Lefèbvre, Gordon D Murray, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the proactive study (prospective pioglitazone clinical trial in macrovascular events): a randomised controlled trial. *The Lancet*, 366(9493):1279–1289, 2005.

[178] Peter Rossing, Philip Hougaard, Knut Borch-Johnsen, and Hans-Henrik Parving. Predictors of mortality in insulin dependent diabetes: 10 year observational follow up study. *Bmj*, 313(7060):779–784, 1996.

[179] Elif I Ekinci, Sophie Clarke, Merlin C Thomas, John L Moran, Karey Cheong, Richard J MacIsaac, and George Jerums. Dietary salt intake and mortality in patients with type 2 diabetes. *Diabetes care*, 34(3):703–709, 2011.

[180] Niels Keiding, Per Kragh Andersen, and John P Klein. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in medicine*, 16(2):215–224, 1997.

[181] Luciano Babuin, Vlad C Vasile, Jose A Rio Perez, Jorge R Alegria, High-Seng Chai, Bekele Afessa, and Allan S Jaffe. Elevated cardiac troponin is an independent risk factor for short-and long-term mortality in medical intensive care unit patients. *Critical care medicine*, 36(3):759–765, 2008.

[182] Peter WF Wilson, Samuel R Bozeman, Tanya M Burton, David C Hoaglin, Rami Ben-Joseph, and Chris L Pashos. Prediction of first events of coronary heart disease and stroke with consideration of adiposity. *Circulation*, 118(2):124–130, 2008.

[183] Bhanukiran Vinzamuri, Yan Li, and Chandan K Reddy. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 241–250. ACM, 2014.

[184] Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5):561–570, 2000.

[185] Andrew P Yu, Yanni F Yu, and Michael B Nichol. Estimating the effect of medication adherence on health outcomes among patients with type 2 diabetesan application of marginal structural models. *Value in Health*, 13(8):1038–1045, 2010.

[186] Arijit Nandi, M Maria Glymour, Ichiro Kawachi, and Tyler J VanderWeele. Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes, and stroke. *Epidemiology (Cambridge, Mass.)*, 23(2):223, 2012.

[187] Aileen JF King. The use of animal models in diabetes research. *British journal of pharmacology*, 166(3):877–894, 2012.

[188] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[189] Kazue Yamaoka and Toshiro Tango. Efficacy of lifestyle education to prevent type 2 diabetes a meta-analysis of randomized controlled trials. *Diabetes care*, 28(11):2780–2786, 2005.

[190] Briana Mezuk, William W Eaton, Sandra Albrecht, and Sherita Hill Golden. Depression and type 2 diabetes over the lifespan a meta-analysis. *Diabetes care*, 31(12):2383–2390, 2008.

[191] Arie Nouwen, K Winkley, J Twisk, CE Lloyd, M Peyrot, K Ismail, F Pouwer, European Depression in Diabetes (EDID) Research Consortium, et al. Type 2 diabetes mellitus as a risk factor for the onset of depression: a systematic review and meta-analysis. *Diabetologia*, 53(12):2480–2486, 2010.

[192] Deborah N Peikes, Lorenzo Moreno, and Sean Michael Orzol. Propensity score matching. *The American Statistician*, 62(3), 2008.

[193] Betty Tao, Massimo Pietropaolo, Mark Atkinson, Desmond Schatz, and David Taylor. Estimating the cost of type 1 diabetes in the us: a propensity score matching method. *PLoS One*, 5(7):e11501, 2010.

[194] Peter C Austin. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, 134(5):1128–1135, 2007.

[195] Kevan R Polkinghorne, Stephen P McDonald, Robert C Atkins, and Peter G Kerr. Vascular access and all-cause mortality: a propensity score analysis. *Journal of the American Society of Nephrology*, 15(2):477–486, 2004.

[196] Hideo Yasunaga, Hiromasa Horiguchi, Kazuaki Kuwabara, Shinya Matsuda, Kiyohide Fushimi, Hideki Hashimoto, and John Z Ayanian. Outcomes after laparoscopic or open distal gastrectomy for early-stage gastric cancer: a propensity-matched analysis. *Annals of surgery*, 257(4):640–646, 2013.

[197] Philip M Short, Samuel IW Lipworth, Douglas HJ Elder, Stuart Schembri, Brian J Lipworth, et al. Effect of β blockers in treatment of chronic obstructive pulmonary disease: a retrospective cohort study. *Bmj*, 342:d2549, 2011.

[198] Oliver Kuss, Benita von Salviati, and Jochen Börgermann. Off-pump versus on-pump coronary artery bypass grafting: A systematic review and meta-analysis of propensity score analyses. *The Journal of Thoracic and Cardiovascular Surgery*, 140(4):829–835, 2010.

[199] Joseph W Hogan and Tony Lancaster. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13(1):17–48, 2004.

[200] Charles H Hennekens, Julie E Buring, and Sherry L Mayrent. *Epidemiology in medicine*. Boston: Little Brown and Company, 1987, 1987.

[201] John M Last, International Epidemiological Association, et al. *A dictionary of epidemiology*, volume 141. Oxford Univ Press, 2001.

[202] David L Sackett, Jonathan J Deeks, and Doughs G Altman. Down with odds ratios! *Evidence based medicine*, 1(6):164–166, 1996.

[203] Jonathan AC Sterne and George Davey Smith. Sifting the evidencewhat's wrong with significance tests? *Physical Therapy*, 81(8):1464–1469, 2001.

[204] David A Grimes and Kenneth F Schulz. An overview of clinical research: the lay of the land. *The lancet*, 359(9300):57–61, 2002.

[205] Edmund F Funai, Emily J Rosenbush, M-J Lee, and Giuseppe Del Priore. Distribution of study designs in four major us journals of obstetrics and gynecology. *Gynecologic and obstetric investigation*, 51(1):8–11, 2001.

[206] Jennifer L Kelsey. *Methods in observational epidemiology*, volume 26. Oxford University Press, USA, 1996.

[207] KJ Rothman and S Greenland. Modern epidemiology.boston: Little brown and company. *Study design for comparison of orthoses*, 273:101–106, 1986.

[208] Richard Doll, Richard Peto, Jillian Boreham, Isabelle Sutherland, et al. Smoking and dementia in male british doctors: prospective study. *Bmj*, 320(7242):1097–1102, 2000.

[209] Philip C Hannaford and CLIFFORD R Kay. The risk of serious illness among oral contraceptive users: evidence from the rcgp's oral contraceptive study. *British journal of general practice*, 48(435):1657–1662, 1998.

[210] Kyung S Kim, Willis L Owen, Deborah Williams, and Lucile L Adams-Campbell. A comparison between bmi and conicity index on predicting coronary heart disease: the framingham heart study. *Annals of epidemiology*, 10(7):424–431, 2000.

[211] Zhiping Huang, Walter C Willett, Graham A Colditz, David J Hunter, JoAnn E Manson, Bernard Rosner, Frank E Speizer, and Susan E Hankinson. Waist circumference, waist: hip ratio, and risk of breast cancer in the nurses' health study. *American journal of epidemiology*, 150(12):1316–1324, 1999.

[212] E Fuller Torrey, R Rawlings, and RH Yolken. The antecedents of psychoses: a case–control study of selected risk factors. *Schizophrenia research*, 46(1):17–23, 2000.

[213] Peng XU, Qiang HUANG, Chen-hai LIU, Fang XIE, Feng SHAO, Cheng-lin ZHU, and Lei LIU. Risk factors for pancreatic cancer: a case-control study. *Tumor*, 31(7):653–657, 2011.

[214] Yoneatsu Osaki and Masumi Minowa. Factors associated with earthquake deaths in the great hanshin-awaji earthquake, 1995. *American journal of epidemiology*, 153(2):153–156, 2001.

[215] Benjamin Avidan, Amnon Sonnenberg, Thomas G Schnell, and Stephen J Sontag. Risk factors for erosive reflux esophagitis: a case-control study. *The American journal of gastroenterology*, 96(1):41–46, 2001.

[216] LAMBERTINA WJ FRENI-TITULAER, DELLE B KELLEY, ADAM G GROW, THOMAS W McKINLEY, FRANK C ARNETT, and MARC C HOCHBERG. Connective tissue disease in southeastern georgia: a case-control study of etiologic factors. *American journal of epidemiology*, 130(2):404–409, 1989.

[217] David E Lilienfeld and Paul D Stolley. *Foundations of epidemiology*. Oxford University Press, USA, 1994.

[218] Valerie Beral, Carol Hermon, Clifford Kay, Philip Hannaford, Sarah Darby, and Gillian Reeves. Mortality associated with oral contraceptive use: 25 year follow up of cohort of 46 000 women from royal college of general practitioners' oral contraception study. *Bmj*, 318(7176):96–100, 1999.

[219] Leo J Seman, Carl DeLuca, Jennifer L Jenner, L Adrienne Cupples, Judith R McNamara, Peter WF Wilson, William P Castelli, Jose M Ordovas, and Ernst J Schaefer. Lipoprotein (a)-cholesterol and coronary heart disease in the framingham heart study. *Clinical chemistry*, 45(7):1039–1046, 1999.

[220] Graham A Colditz, Bernard A Rosner, Frank E Speizer, Nurses' Health Study Research Group, et al. Risk factors for breast cancer according to family history of breast cancer. *Journal of the National Cancer Institute*, 88(6):365–371, 1996.

[221] David L Sackett. Bias in analytic research. *Journal of chronic diseases*, 32(1):51–63, 1979.

[222] James Lee. Odds ratio or relative risk for cross-sectional data? *International Journal of Epidemiology*, 23(1):201–203, 1994.

[223] CJ Mann. Observational research methods. research design ii: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, 20(1):54–60, 2003.

[224] Jeffrey V Johnson and Ellen M Hall. Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the swedish working population. *American journal of public health*, 78(10):1336–1342, 1988.

[225] Lars Bo Andersen, Maarike Harro, Luis B Sardinha, Karsten Froberg, Ulf Ekelund, Søren Brage, and Sigmund Alfred Anderssen. Physical activity and clustered cardiovascular risk in children: a cross-sectional study (the european youth heart study). *The Lancet*, 368(9532):299–304, 2006.

[226] CE Coffey, WE Wilkinson, LA Parashos, SAR Soady, RJ Sullivan, LJ Patterson, GS Figiel, MC Webb, CE Spritzer, and WT Djang. Quantitative cerebral anatomy of the aging human brain a cross-sectional study using magnetic resonance imaging. *Neurology*, 42(3):527–527, 1992.

[227] Kate Ann Levin. Study design iii: Cross-sectional studies. *Evidence-based dentistry*, 7(1):24–25, 2006.

[228] Thomas Gabert and Dean T Stueland. Recreational injuries and deaths in northern wisconsin: analysis of injuries and fatalities from snowmobiles over 3 years. *Wisconsin medical journal*, 92(12):671–675, 1993.

[229] BD Krane, Maria Angela Ricci, WB Sweeney, and NARAYAN Deshmukh. All-terrain vehicle injuries. a review at a rural level ii trauma center. *The American Surgeon*, 54(8):471–474, 1988.

[230] B Jaremin, E Kotulak, M Starnawska, W Mroziński, and E Wojciechowski. Death at sea: certain factors responsible for occupational hazard in polish seamen and deep-sea fishermen. *International journal of occupational medicine and environmental health*, 10(4):405–416, 1996.

[231] Lynn M Marshall, Donna Spiegelman, Robert L Barbieri, Marlene B Goldman, JoAnn E Manson, Graham A Colditz, Walter C Willett, and David J Hunter. Variation in the incidence of uterine leiomyoma among premenopausal women by age and race. *Obstetrics & Gynecology*, 90(6):967–973, 1997.

[232] Sharon H Giordano, Aman U Buzdar, and Gabriel N Hortobagyi. Breast cancer in men. *Annals of internal medicine*, 137(8):678–687, 2002.

[233] Joli R Weiss, Kirsten B Moysich, and Helen Swede. Epidemiology of male breast cancer. *Cancer Epidemiology Biomarkers & Prevention*, 14(1):20–26, 2005.

[234] Frederick A Anderson, H Brownell Wheeler, Robert J Goldberg, David W Hosmer, Nilima A Patwardhan, Borko Jovanovic, Ann Forcier, and James E Dalen. A population-based perspective of the hospital incidence and case-fatality rates of deep vein thrombosis and pulmonary embolism: the worcester dvt study. *Archives of internal medicine*, 151(5):933–938, 1991.

[235] Suzanne C Tough, Calvin A Greene, Larry W Svenson, and Jaques Belik. Effects of in vitro fertilization on low birth weight, preterm delivery, and multiple birth. *The Journal of pediatrics*, 136(5):618–622, 2000.

[236] D Bider, A Livshitz, I Tur Kaspa, A Shulman, J Levron, and J Dor. Incidence and perinatal outcome of multiple pregnancies after intracytoplasmic sperm injection compared to standard in vitro fertilization. *Journal of assisted reproduction and genetics*, 16(5):221–226, 1999.

[237] Alistair Dunn and Alison Macfarlane. Recent trends in the incidence of multiple births and associated mortality in england and wales. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 75(1):F10–F19, 1996.

[238] RP Steegers-Theunissen, Werner M Zwertbroek, AJ Huisjes, Humphrey H Kanhai, Hein W Bruinse, and HM Merkus. Multiple birth prevalence in the netherlands. impact of maternal age and assisted reproductive techniques. *The Journal of reproductive medicine*, 43(3):173–179, 1998.

[239] David Gotz, Jimeng Sun, Nan Cao, and Shahram Ebadollahi. Visual cluster analysis in support of clinical decision intelligence. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:481–90, January 2011.

[240] Anna Bauer-Mehren, Paea Lependu, Srinivasan V Iyer, Rave Harpaz, Nicholas J Leeper, and Nigam H Shah. Network analysis of unstructured EHR data for clinical research. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2013:14–8, January 2013.

[241] Leila Kalankesh, James Weatherall, Thamer Ba-Dhfari, Iain Buchan, and Andy Brass. Taming EHR data: using semantic similarity to reduce dimensionality. *Studies in health technology and informatics*, 192:52–6, January 2013.

[242] George H Dunteman. *Principal components analysis*. Number 69. Sage, 1989.

[243] Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.

[244] Chengqi Zhang and Shichao Zhang. *Association rule mining: models and algorithms*. Springer-Verlag, 2002.

[245] Adam Wright, Elizabeth S Chen, and Francine L Maloney. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of biomedical informatics*, 43(6):891–901, December 2010.

[246] David A Hanauer, Daniel R Rhodes, and Arul M Chinnaiyan. Exploring clinical associations using '-omics' based enrichment analyses. *PloS one*, 4(4):e5203, January 2009.

[247] David A Hanauer and Naren Ramakrishnan. Modeling temporal relationships in large scale clinical associations. *Journal of the American Medical Informatics Association : JAMIA*, 20(2):332–41.

[248] Vance Liao and Ming-Syan Chen. Efficient mining gapped sequential patterns for motifs in biological sequences. *BMC systems biology*, 7 Suppl 4:S7, January 2013.

[249] George Hripcsak and David J Albers. Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association : JAMIA*, 20(e2):e311–8, December 2013.

[250] Y Shahar. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90(1-2):79–133, February 1997.

[251] Lucia Sacchi, Cristiana Larizza, Carlo Combi, and Riccardo Bellazzi. Data mining with Temporal Abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*, 15(2):217–247, June 2007.

[252] Huidong Warren Jin, Jie Chen, Hongxing He, Graham J Williams, Chris Kelman, and Christine M O'Keefe. Mining unexpected temporal associations: applications in detecting adverse drug reactions. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 12(4):488–500, July 2008.

[253] Iyad Batal, Lucia Sacchi, and Riccardo Bellazzi. Multivariate Time Series Classification with Temporal Abstractions. 1994.

[254] Shima Ghassempour, Federico Girosi, and Anthony Maeder. Clustering multivariate time series using hidden markov models. *International journal of environmental research and public health*, 11(3):2741–2763, 2014.

[255] Ying Li, Hojjat Salmasian, Santiago Vilar, Herbert Chase, Carol Friedman, and Ying Wei. A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 21(2):308–14.

[256] Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association : JAMIA*, 20(e2):e206–11, December 2013.

[257] Robert J Carroll, Anne E Eyler, and Joshua C Denny. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:189–96, January 2011.

[258] Hua Xu, Zhenming Fu, Anushi Shah, Yukun Chen, Neeraja B Peterson, Qingxia Chen, Subramani Mani, Mia A Levy, Qi Dai, and Josh C Denny. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:1564–72, January 2011.

[259] Iyad Batal and Milos Hauskrecht. Mining Clinical Data using Minimal Predictive Rules. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:31–35, January 2010.

[260] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, and Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, page 280, 2012.

[261] Peter J F Lucas, Linda C van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine*, 30(3):201–14, March 2004.

[262] Yuriy Sverchkov, Shyam Visweswaran, Gilles Clermont, Milos Hauskrecht, and Gregory F Cooper. A multivariate probabilistic method for comparing two clinical datasets. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 795–800. ACM, 2012.

[263] Subramani Mani, Asli Ozdas, Constantin Aliferis, Huseyin Atakan Varol, Qingxia Chen, Randy Carnevale, Yukun Chen, Joann Romano-Keeler, Hui Nian, and Jörn-Hendrik Weitkamp. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. *Journal of the American Medical Informatics Association : JAMIA*, 21(2):326–36.

[264] Chayan Sarkar and Jaideep Srivastava. Impact of density of lab data in ehr for prediction of potentially preventable events. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 529–534. IEEE, 2013.

[265] Sharath Cholleti, Andrew Post, Jingjing Gao, Xia Lin, William Bornstein, Dedra Cantrell, and Joel Saltz. Leveraging derived data elements in data analytic models for understanding and predicting hospital readmissions. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:103–11, January 2012.

[266] Yubin Park and Joydeep Ghosh. A hierarchical ensemble of α-trees for predicting expensive hospital visits. In *Brain Informatics and Health*, pages 178–187. Springer, 2014.

[267] Emily Kawaler, Alexander Cobian, Peggy Peissig, Deanna Cross, Steve Yale, and Mark Craven. Learning to predict post-hospitalization VTE risk from EHR data. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:436–45, January 2012.

[268] Haijun Zhai, Patrick Brady, Qi Li, Todd Lingren, Yizhao Ni, Derek S Wheeler, and Imre Solti. Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. *Resuscitation*, 85(8):1065–71, August 2014.

[269] Subramani Mani, Yukun Chen, Tom Elasy, Warren Clayton, and Joshua Denny. Type 2 diabetes risk forecasting from EMR data using machine learning. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:606–15, January 2012.

[270] Shreyas Karnik, Sin Lam Tan, Bess Berg, Ingrid Glurich, Jinfeng Zhang, Humberto J Vidaillet, C David Page, and Rajesh Chowdhary. Predicting atrial fibrillation and flutter using electronic health records. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2012:5562–5, January 2012.

[271] Jason Scott Mathias, Ankit Agrawal, Joe Feinglass, Andrew J Cooper, David William Baker, and Alok Choudhary. 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 : JAMIA*, 20(e1):e118–24, June 2013.

[272] Sandy H Huang, Paea LePendu, Srinivasan V Iyer, Ming Tai-Seale, David Carrell, and Nigam H Shah. Toward personalizing treatment for depression: predicting diagnosis and severity. *Journal of the American Medical Informatics Association : JAMIA*, July 2014.

[273] Asil Oztekin, Dursun Delen, and Zhenyu James Kong. Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. *International journal of medical informatics*, 78(12):e84–96, December 2009.

[274] A. Soltani Sarvestani, A. A. Safavi, N.M. Parandeh, and M. Salehi. Predicting breast cancer survivability using data mining techniques. In *2010 2nd International Conference on Software Technology and Engineering*, volume 2, pages V2–227–V2–231. IEEE, October 2010.

[275] Marios Skevofilakas, Konstantia Zarkogianni, Basil G Karamanos, and Konstantina S Nikita. A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2010:6713–6, January 2010.

[276] 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. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Medical decision making : an international journal of the Society for Medical Decision Making*, 30(6):639–50.

[277] Vitaly Herasevich, Daryl J Kor, Arun Subramanian, and Brian W Pickering. Connecting the dots: rule-based decision support systems in the modern EMR era. *Journal of clinical monitoring and computing*, 27(4):443–8, August 2013.

[278] João Maroco, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabel Santana, and Alexandre de Mendonça. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes*, 4:299, January 2011.

[279] Peter C Austin, Douglas S Lee, Ewout W Steyerberg, and Jack V Tu. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? *Biometrical journal. Biometrische Zeitschrift*, 54(5):657–673, September 2012.

[280] Ying-Jui Chang, Min-Li Yeh, Yu-Chuan Li, Chien-Yeh Hsu, Chao-Cheng Lin, Meng-Shiuan Hsu, and Wen-Ta Chiu. Predicting hospital-acquired infections by scoring system with simple parameters. *PloS one*, 6(8):e23137, January 2011.

[281] Yuri Ito, Akiko Ioka, Hideaki Tsukuma, Wakiko Ajiki, Tomoyuki Sugimoto, Bernard Rachet, and Michel P Coleman. Regional differences in population-based cancer survival between six prefectures in Japan: application of relative survival models with funnel plots. *Cancer science*, 100(7):1306–11, July 2009.

[282] Michel P Coleman, Manuela Quaresma, Franco Berrino, Jean-Michel Lutz, Roberta De Angelis, Riccardo Capocaccia, Paolo Baili, Bernard Rachet, Gemma Gatta, Timo Hakulinen, Andrea Micheli, Milena Sant, Hannah K Weir, J Mark Elwood, Hideaki Tsukuma, Sergio Koifman, Gulnar Azevedo E Silva, Silvia Francisci, Mariano Santaquilani, Arduino Verdecchia, Hans H Storm, and John L Young. Cancer survival in five continents: a worldwide population-based study (CONCORD). *The lancet oncology*, 9(8):730–56, August 2008.

[283] Etsuko Nomura, Hideaki Tsukuma, Wakiko Ajiki, and Akira Oshima. Population-based study of relationship between hospital surgical volume and 5-year survival of stomach cancer patients in Osaka, Japan. *Cancer science*, 94(11):998–1002, November 2003.

[284] Andrea H Ramirez, Yaping Shi, Jonathan S Schildcrout, Jessica T Delaney, Hua Xu, Matthew T Oetjens, Rebecca L Zuvich, Melissa A Basford, Erica Bowton, Min Jiang, Peter Speltz, Raquel Zink, James Cowan, Jill M Pulley, Marylyn D Ritchie, Daniel R Masys, Dan M Roden,

Dana C Crawford, and Joshua C Denny. Predicting warfarin dosage in European-Americans and African-Americans using DNA samples linked to an electronic health record. *Pharmacogenomics*, 13(4):407–18, March 2012.

[285] Deepika Goyal, Elsie J Wang, Jeremy Shen, Eric C Wong, and Latha P Palaniappan. Clinically identified postpartum depression in Asian American mothers. *Journal of obstetric, gynecologic, and neonatal nursing : JOGNN / NAACOG*, 41(3):408–16.

[286] Mohamed Ghalwash and Zoran Obradovic. A data-driven model for optimizing therapy duration for septic patients. In *Proc. 14th SIAM Intl. Conf. Data Mining, 3rd Workshop on Data Mining for Medicine and Healthcare, Philadelphia, PA, USA (April 2014)*.

[287] Jimeng Sun, Candace D McNaughton, Ping Zhang, Adam Perer, Aris Gkoulalas-Divanis, Joshua C Denny, Jacqueline Kirby, Thomas Lasko, Alexander Saip, and Bradley A Malin. Predicting changes in hypertension control using electronic health records from a chronic disease management program. *Journal of the American Medical Informatics Association*, 21(2):337–344, 2014.

[288] Zheng Wang, Prithwish Chakraborty, Sumiko R Mekaru, John S Brownstein, Jieping Ye, and Naren Ramakrishnan. Dynamic poisson autoregression for influenza-like-illness case count prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294. ACM, 2015.

[289] M Panahiazar, V Taslimitehrani, N Pereira, and J Pathak. Using ehrs and machine learning for heart failure survival analysis. In *MEDINFO 2015: EHealth-enabled Health: Proceedings of the 15th World Congress on Health and Biomedical Informatics*, volume 216, page 40. IOS Press, 2015.

[290] Fei Wang, Ping Zhang, Buyue Qian, Xiang Wang, and Ian Davidson. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154. ACM, 2014.

[291] Brian J Wells, Anil Jain, Susana Arrigain, Changhong Yu, Wayne A Rosenkrans, and Michael W Kattan. Predicting 6-year mortality risk in patients with type 2 diabetes. *Diabetes Care*, 31(12):2301–2306, 2008.

[292] Bhanukiran Vinzamuri and Chandan K Reddy. Cox regression with correlation based regularization for electronic health records. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 757–766. IEEE, 2013.

[293] Sunil J Rao. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. *Journal of the American Statistical Association*, 98(461):257–258, 2003.

[294] Jeffrey M Wooldridge. Some alternatives to the box-cox regression model. *International Economic Review*, pages 935–955, 1992.

[295] Kenji Ikeda, Hiromitsu Kumada, Satoshi Saitoh, Yasuji Arase, and Kazuaki Chayama. Effect of repeated transcatheter arterial embolization on the survival time in patients with hepatocellular carcinoma. an analysis by the cox proportional hazard model. *Cancer*, 68(10):2150–2154, 1991.

[296] Kung-Yee Liang, Steven G Self, and Xinhua Liu. The cox proportional hazards model with change point: An epidemiologic application. *Biometrics*, pages 783–793, 1990.

[297] Thomas Lumley, Richard A Kronmal, Mary Cushman, Teri A Manolio, and Steven Goldstein. A stroke prediction score in the elderly: validation and web-based application. *Journal of clinical epidemiology*, 55(2):129–136, 2002.

[298] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. A hazard based approach to user return time prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1719–1728. ACM, 2014.

[299] Robert Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[300] Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, et al. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13, 2011.

[301] Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

[302] Terry Therneau and Cindy Crowson. Using time dependent covariates and time dependent coefficients in the cox model. *Red*, 2:1, 2014.

[303] Shivapratap Gopakumar, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Stabilizing sparse cox model using clinical structures in electronic medical records. *arXiv preprint arXiv:1407.6094*, 2014.

[304] Wei Zhang, Takayo Ota, Viji Shridhar, Jeremy Chien, Baolin Wu, and Rui Kuang. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS computational biology*, 9(3):e1002975, 2013.

[305] Marti A. Hearst, Susan T Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[306] Faisal M Khan and Valentina Bayer Zubek. Support vector regression for censored data (svrc): a novel tool for survival analysis. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 863–868. IEEE, 2008.

[307] Ludger Evers and Claudia-Martina Messow. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, 24(14):1632–1638, 2008.

[308] Pannagadatta K Shivaswamy, Wei Chu, and Martin Jansche. A support vector approach to censored targets. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 655–660. IEEE, 2007.

[309] Vanya Van Belle, Kristiaan Pelckmans, JAK Suykens, and Sabine Van Huffel. Support vector machines for survival analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, pages 1–8, 2007.

[310] Han-Tai Shiao and Vladimir Cherkassky. Learning using privileged information (lupi) for modeling survival data. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 1042–1049. IEEE, 2014.

[311] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, and Honglak Lee. An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 183–192. ACM, 2010.

[312] L Gordon and RA Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69(10):1065–1069, 1985.

[313] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

[314] Michael W Kattan, Kenneth R Hess, and J Robert Beck. Experiments to determine whether recursive partitioning (cart) or an artificial neural network overcomes theoretical limitations of cox proportional hazards regression. *Computers and biomedical research*, 31(5):363–373, 1998.

[315] Peter B Snow, Deborah S Smith, and William J Catalona. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. *The Journal of urology*, 152(5 Pt 2):1923–1926, 1994.

[316]

[317] Santu Rana, Sunil Gupta, Dinh Phung, and Svetha Venkatesh. A predictive framework for modeling healthcare data with evolving clinical interventions. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(3):162–182, 2015.

[318] Micol Sandri, Paola Berchialla, Ileana Baldi, Dario Gregori, and Roberto Alberto De Blasi. Dynamic Bayesian Networks to predict sequences of organ failures in patients admitted to ICU. *Journal of biomedical informatics*, 48:106–13, April 2014.

[319] E. Gatti, D. Luciani, and F. Stella. A continuous time Bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*, 24(4):496–515, December 2011.

[320] Linda Peelen, Nicolette F de Keizer, Evert de Jonge, Robert-Jan Bosman, Ameen Abu-Hanna, and Niels Peek. Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit. *Journal of biomedical informatics*, 43(2):273–86, April 2010.

[321] Dipankar Sengupta and Pradeep K Naik. SN algorithm: analysis of temporal clinical data for mining periodic patterns and impending augury. *Journal of clinical bioinformatics*, 3(1):24, January 2013.

[322] Marion Verduijn, Lucia Sacchi, Niels Peek, Riccardo Bellazzi, Evert de Jonge, and Bas A J M de Mol. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. *Artificial intelligence in medicine*, 41(1):1–12, September 2007.

[323] Girish N Nadkarni, Omri Gottesman, James G Linneman, Herbert Chase, Richard L Berg, Samira Farouk, Rajiv Nadukuru, Vaneet Lotay, Steve Ellis, George Hripcsak, et al. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. In *AMIA Annual Symposium Proceedings*, volume 2014, page 907. American Medical Informatics Association, 2014.

[324] David J Albers, Noémie Elhadad, E Tabak, Adler Perotte, and George Hripcsak. Dynamical phenotyping: using temporal analysis of clinically collected physiologic data to stratify populations. 2014.

[325] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):103, January 2011.

[326] Gyorgy J. Simon, Vipin Kumar, and Peter W. Li. A simple statistical model and association rule filtering for classification. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, page 823, New York, New York, USA, August 2011. ACM Press.

[327] Serguei V S Pakhomov, Nilay D Shah, Holly K Van Houten, Penny L Hanson, and Steven A Smith. The role of the electronic medical record in the assessment of health related quality of life. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:1080–8, January 2011.

[328] Glenn Fung, Shipeng Yu, Cary Dehing-Oberije, Dirk De Ruysscher, Philippe Lambin, Sriram Krishnan, and R Rao Bharat. Privacy-preserving predictive models for lung cancer survival analysis. *Practical Privacy-Preserving Data Mining*, page 40, 2008.

[329] Maarten van der Heijden, Marina Velikova, and Peter J F Lucas. Learning Bayesian networks for clinical time series analysis. *Journal of biomedical informatics*, 48:94–105, April 2014.

[330] Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2013:1109–15, January 2013.

[331] Joseph L Breault, Colin R Goodall, and Peter J Fos. Data mining a diabetic data warehouse. *Artificial intelligence in medicine*, 26(1-2):37–54.

[332] Bonnie Westra, Sanjoy Dey, Gang Fang, Michael Steinbach, Vipin Kumar, Cristina Oancea, Kay Savik, and Mary Dierich. Interpretable predictive models for knowledge discovery from home-care electronic health records. *Journal of Healthcare Engineering*, 2(1):55–74, 2011.

[333] George Mathew and Zoran Obradovic. Distributed privacy-preserving decision support system for highly imbalanced clinical data. *ACM Transactions on Management Information Systems (TMIS)*, 4(3):12, 2013.

[334] Yubin Park and Joydeep Ghosh. Pegs: Perturbed gibbs samplers that gen-erate privacy-compliant synthetic data.

[335] Nitesh V Chawla and Darcy A Davis. Bringing big data to personalized healthcare: a patient-centered framework. *Journal of general internal medicine*, 28(3):660–665, 2013.

[336] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun. Publishing data from electronic health records while preserving privacy: A survey of algorithms. *Journal of biomedical informatics*, 50:4–19, 2014.

[337] A Gkoulalas-Divanis, Grigorios Loukides, and Jian Sun. Toward smarter healthcare: Anonymizing medical data to support research studies. *IBM Journal of Research and Development*, 58(1):9–1, 2014.

[338] T Ryan Hoens, Marina Blanton, Aaron Steele, and Nitesh V Chawla. Reliable medical recommendation systems with patient privacy. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4):67, 2013.

[339] Jesus J Caban and David Gotz. Visual analytics in healthcare–opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262, 2015.

[340] Zhiyuan Zhang, David Gotz, and Adam Perer. Iterative cohort analysis and exploration. *Information Visualization*, page 1473871614526077, 2014.

[341] David Gotz, Harry Stavropoulos, Jimeng Sun, and Fei Wang. Icda: a platform for intelligent care delivery analytics. In *AMIA annual symposium proceedings*, volume 2012, page 264. American Medical Informatics Association, 2012.

[342] DH Gotz, Jian Sun, and Nianxia Cao. Multifaceted visual analytics for healthcare applications. *IBM Journal of Research and Development*, 56(5):6–1, 2012.

[343] Jimeng Sun, David Gotz, and Nan Cao. Diseaseatlas: Multi-facet visual analytics for online disease articles. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 1123–1126. IEEE, 2010.

[344] Nan Cao, David Gotz, Jimeng Sun, Yu-Ru Lin, and Huamin Qu. Solarmap: multifaceted visual analytics for topic exploration. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 101–110. IEEE, 2011.

[345] David Gotz, Jimeng Sun, Nan Cao, and Shahram Ebadollahi. Visual cluster analysis in support of clinical decision intelligence. In *AMIA Annual Symposium Proceedings*, volume 2011, page 481. American Medical Informatics Association, 2011.

[346] Nan Cao, David Gotz, Jimeng Sun, and Huamin Qu. Dicon: Interactive visual analysis of multidimensional clusters. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2581–2590, 2011.

[347] Adam Perer and Jimeng Sun. Matrixflow: temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, volume 2012, page 716. American Medical Informatics Association, 2012.

[348] Adam Perer and Fei Wang. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 153–162. ACM, 2014.

[349] Jimeng Sun, Daby Sow, Jianying Hu, and Shahram Ebadollahi. Localized supervised metric learning on temporal physiological data. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4149–4152. IEEE, 2010.

[350] Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu, and Shahram Ebadollahi. Sor: scalable orthogonal regression for non-redundant feature selection and its healthcare applications. In *SIAM data mining conference*, pages 576–587. SIAM, 2012.

[351] Jiayu Zhou, Zhaosong Lu, Jimeng Sun, Lei Yuan, Fei Wang, and Jieping Ye. Feafiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1034–1042. ACM, 2013.

[352] Jiayu Zhou, Jimeng Sun, Yashu Liu, Jianying Hu, and Jieping Ye. Patient risk prediction model via top-k stability selection. In *SIAM Conference on Data Mining*. SIAM, 2013.

[353] Robert N Anderson, National Center for Health Statistics (US), et al. Deaths: leading causes for 2001. 2003.

[354] Lars Rydén, Eberhard Standl, Małgorzata Bartnik, Greet Van den Berghe, John Betteridge, Menko-Jan De Boer, Francesco Cosentino, Bengt Jönsson, Markku Laakso, Klas Malmberg, et al. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: executive summary. *European heart journal*, 28(1):88–136, 2007.

[355] Linda E Lévesque, James A Hanley, Abbas Kezouh, and Samy Suissa. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *Bmj*, 340:b5087, 2010.

[356] A Elixhauser, C Steiner, and L Palmer. Clinical classifications software (ccs). *Book Clinical Classifications Software (CCS)(Editor ed^ eds)*, 2008.

[357] Dingcheng Li, Gyorgy Simon, Christopher G Chute, and Jyotishman Pathak. Using association rule mining for phenotype extraction from electronic health records. *AMIA Summits on Translational Science Proceedings*, 2013:142, 2013.

[358] Nicholas J Wareham and Stephen ORahilly. The changing classification and diagnosis of diabetes: new classification is based on pathogenesis, not insulin dependence. *BMJ: British Medical Journal*, 317(7155):359, 1998.

[359] Mike Conway, Richard L Berg, David Carrell, Joshua C Denny, Abel N Kho, Iftikhar J Kullo, James G Linneman, Jennifer A Pacheco, Peggy Peissig, Luke Rasmussen, et al. Analyzing the heterogeneity and complexity of electronic health record oriented phenotyping algorithms. In *AMIA annual symposium proceedings*, volume 2011, page 274. American Medical Informatics Association, 2011.

[360] Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2):e253–e259, 2013.

[361] Peter WF Wilson, James B Meigs, Lisa Sullivan, Caroline S Fox, David M Nathan, and Ralph B DAgostino. Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study. *Archives of internal medicine*, 167(10):1068–1074, 2007.

[362] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):103, 2011.

[363] Gyorgy J Simon, John Schrom, M Regina Castro, Peter W Li, and Pedro J Caraballo. Survival association rule mining towards type 2 diabetes risk assessment. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1293. American Medical Informatics Association, 2013.

[364] Ana F Macedo, Fiona C Taylor, Juan P Casas, Alma Adler, David Prieto-Merino, and Shah Ebrahim. Unintended effects of statins from observational studies in the general population: systematic review and meta-analysis. *BMC medicine*, 12(1):51, 2014.

[365] Fiona Taylor, Mark D Huffman, Ana Filipa Macedo, TH Moore, Margaret Burke, George Davey Smith, Kirsten Ward, and Shah Ebrahim. Statins for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev*, 1(1), 2013.

[366] Swapnil N Rajpathak, Dharam J Kumbhani, Jill Crandall, Nir Barzilai, Michael Alderman, and Paul M Ridker. Statin therapy and risk of developing type 2 diabetes: a meta-analysis. *Diabetes care*, 32(10):1924–1929, 2009.

[367] John R Schrom, Pedro J Caraballo, M Regina Castro, and György J Simon. Quantifying the effect of statin use in pre-diabetic phenotypes discovered through association rule mining. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1249. American Medical Informatics Association, 2013.